Artificial Intelligence, speech and language processing approaches to monitoring Alzheimer's Disease: a systematic review

Artificial Intelligence, speech and language processing approaches to monitoring Alzheimer's Disease: a systematic review

Sofia **de la Fuente Garcia** [a,*], Craig **Ritchie** [b] and Saturnino **Luz** [a]
[a] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, Scotland, UK*
E-mails: sofia.delafuente@ed.ac.uk, s.luz@ed.ac.uk
[b] *Centre for Dementia Prevention, The University of Edinburgh, Scotland, UK*
E-mail: craig.ritchie@ed.ac.uk

Abstract.
Background: Language is a valuable source of clinical information in Alzheimer's Disease, as it declines concurrently with neurodegeneration. Consequently, speech and language data have been extensively studied in connection with its diagnosis.
Objective: firstly, to summarise the existing findings on the use of artificial intelligence, speech and language processing to predict cognitive decline in the context of Alzheimer's Disease. Secondly, to detail current research procedures, highlight their limitations and suggest strategies to address them.
Method: Systematic review of original research between 2000 and 2019, registered in PROSPERO (reference CRD42018116606). An interdisciplinary search covered six databases on engineering (ACM and IEEE), psychology (PsycINFO), medicine (PubMed and Embase) and Web of Science. Bibliographies of relevant papers were screened until December 2019.
Results: from 3,654 search results 51 articles were selected against the eligibility criteria. Four tables summarise their findings: *study details*, (aim, population, interventions, comparisons, methods and outcomes), *data details* (size, type, modalities, annotation, balance, availability and language of study), *methodology* (pre-processing, feature generation, machine learning, evaluation and results) and *clinical applicability* (research implications, clinical potential, risk of bias and strengths/limitations).
Conclusion: promising results are reported across nearly all 51 studies, but very few have been implemented in clinical research or practice. The main limitations of the field are poor standardisation, limited comparability of results, and a degree of disconnect between study aims and clinical applications. Active attempts to close these gaps will support translation of future research into clinical practice.

Keywords: screening, Alzheimer's Disease, dementia, cognitive decline, computational linguistics, speech processing, machine learning, artificial intelligence

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that involves decline of cognitive and functional abilities as the illness progresses [1]. It is the most common aetiology of dementia. Given its prevalence, it has effects beyond just patients and carers as it also has a severe societal and economic impact worldwide [2]. Although memory loss is often considered the signature symptom of AD, language impairment may also appear in its early stages [3]. Consequently, and due to the ubiquitous nature of speech and language, multiple studies rely on these modalities as sources of clinical information for AD, from foundational qualitative research [e.g. 4, 5] to more recent work on computational speech technology [e.g. 6–8]. The potential for using speech as a biomarker

*Corresponding author. E-mail: sofia.delafuente@ed.ac.uk. Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX

for AD is based on several prospective values, including: 1) the ease with which speech can be recorded and tracked over time, 2) its non-invasiveness, 3) the fact that technologies for speech analysis have improved markedly in the past decade, boosted by advances in artificial intelligence (AI) and machine learning, and 4) the fact that speech problems may be manifest at different stages of the disease, making it a life-course assessment that has value unlimited by disease stage.

Recent studies on the use of AI in AD research entail using language and speech data collected in different ways and applying computational speech processing for diagnosis, prognosis or progression modelling. This technology encompasses methods for recognizing, analysing and understanding spoken discourse. It implies that at least part of the AD detection process could be automated (passive). Machine learning methods have been central to this research programme. Machine learning is a field of AI that concerns itself with the induction of predictive models "learnt" directly from data, where the learner improves its own performance through "experience" (i.e. exposure to greater amounts of data). Research on automatic processing of speech and language with AI and machine learning methods have yielded encouraging results and attracted increasing interest. Different approaches have been studied, including computational linguistics [e.g. 9], computational paralinguistics [e.g. 10], signal processing [e.g. 11] and human-robot interaction [e.g. 12].

However, investigations of the use of language and speech technology in Alzheimer's research are heterogeneous, which makes consensus, conclusions and translation into larger studies or clinical practice problematic. The range of goals pursued in such studies is also broad, including automated screening for early Alzheimer's disease, tools for early detection of disease in clinical practice, monitoring of disease progression and signalling potential mechanistic underpinnings to speech problems at a biological level thereby improving disease models. Despite progress in research, the small, inconsistent, single-lab and non-standardised nature of most studies has yielded results that are not robust enough to be aggregated and thereafter implemented towards those goals. This has resulted in gaps between research contexts, clinical potential and actual clinical applications of this new technology.

We sought to summarise the current state of the evidence regarding AI approaches in speech analysis for Alzheimer's disease with a view to setting a foundation for future research in this area and potential development of guidelines for research and implementation. The review has three main objectives. Firstly, to present the main aims and findings of this research, secondly to outline the main methodological approaches and finally surmise the potential for each technique to be ready for further evaluation towards clinical use. In doing so we hope to contribute to the development of these novel, exciting, and yet underutilised approaches, towards clinical practice.

Methods

The procedures adopted in this review were specified in a protocol registered with the international prospective register of systematic reviews PROSPERO (reference: CRD42018116606). In the following sections we describe the elegibility criteria, information sources, search strategy, study records management, study records selection, data collection process, data items (extraction tool), risk of bias in individual studies, data synthesis, meta-bias(es) and confidence in cumulative evidence.

Elegibility criteria

We aimed to summarise all available scientific studies where an interactive artificial intelligence (AI) approach was adopted for neuropsychological monitoring. Interaction-based technology entails data obtained through a form of communication, and AI entails some automation of the process. Therefore, we included articles where automatic machine learning methods were used for AD screening, detection and prediction, by means of computational linguistics and/or speech technology.

Articles were deemed eligible if they described studies of neurodegeneration in the context of AD. That is, subjective cognitive impairment (SCI), mild cognitive impairment (MCI), AD or other dementia-related terminology if indicated as AD-related in the full text (e.g. if a paper title reads unspecified "dementia" but the research field is AD). The included studies examined behavioural patterns that may precede overt cognitive decline as well as observable cognitive impairment in these neurodegenerative diseases. Related conditions such as semantic dementia (a form of aphasia) or Parkinson's Disease (a different neurodegenerative disease) formed part of the exclusion criteria (except when in comorbidity with AD). Language was not

an exclusion criterion, and translation resources were used as appropriate.

Another exclusion criterion is the exclusive use of traditional statistics in the analysis. The inclusion criteria require at least one component of AI, ML or big data, even if the study encompasses traditional statistical analysis. Further exclusion criteria apply to related studies relying exclusively on neuroimaging techniques such as magnetic resonance imaging (MRI), with no relation to language or speech, even if they do implement AI methods. The same applies to biomarker studies (e.g. APOE genotyping). This review also excluded purely epidemiological studies, that is, studies aimed at analysing the distribution of the condition rather than assessing the potential of AI tools for monitoring its progress.

In terms of publication status, we considered peer-reviewed journal and conference articles only. Records that were not original research papers were excluded (i.e. conference abstracts and systematic reviews). In order to avoid redundancy, we assessed research by the same group and excluded overlapping publications. This was assessed by reading the text in full and selecting only the most relevant article for review (i.e. most comprehensive and up to date). Due to limited resources, we also excluded papers when full-texts were found unavailable in all our alternative sources.

Lastly, we considered papers from a twenty-year span, from the beginning of 2000 to the end of 2019, anticipating that the closer to the end of this timeframe, the larger the number of results, as shown in Figure 1.

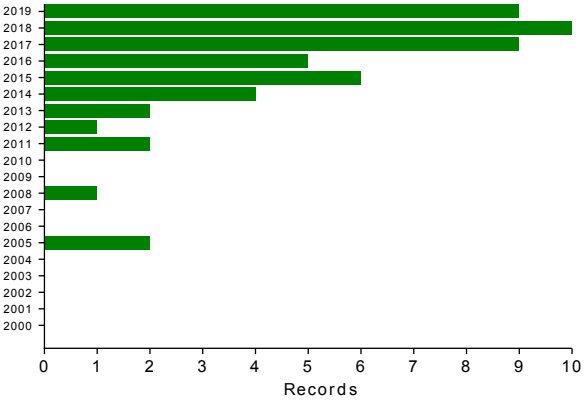

Fig. 1. Number of records found suitable for review each year (2000-2019).

Information Sources

Between October and December 2019, we searched the following electronic databases: ACM, Embase, IEEE, PsycINFO, PubMed, and Web of Science. We contacted study authors by email when full-text versions of relevant papers where not available through the university library, with varying degrees of success.

We also included relevant titles found through "forward citation tracking" with Google Scholar, screening articles references and research portal suggestions suggestions.

Search Strategy

Given the heterogeneity of the field, a broad search needed to be conducted. For the health condition of interest, AD, we included terms such as dementia, cognitive decline and Alzheimer. For the methodology, we included speech, technology, analysis and natural language processing, artificial intelligence, machine learning, and big data.

The search strategy was developed collaboratively between the authors, and with the help of the University of Edinburgh's academic support librarian. After a few iterations and trials, we decided not to include the AI terms, since this seemed to constrain the search too much, yielding fewer results. Therefore, the search queries were specified as follows (example for PubMed):

- (speech AND (dementia OR "cognitive decline" OR (cognit* AND impair*) OR Alzheimer) AND (technology OR analysis)) OR ("natural language processing" AND (dementia OR "cognitive decline" OR (cognit* AND impair*) OR Alzheimer))
- Filters applied: 01/01/2000 - 31/12/2019.

Then, we applied the exclusion criteria, starting from the lack of AI, ML and big data methods, usually detected in the abstract.

We used EndNote X8 [13] for study records management.

Study records selection

Screening for record selection happened in two phases, independently undertaken by two reviewers and following pre-established eligibility criteria. In the first phase, the two independent authors screened titles and abstracts against exclusion criteria using EndNote.

The second phase consisted of a full-text screening for those papers that could not be absolutely included or excluded based on title and abstract information only. Any emerging titles that were deemed relevant were added to the screening process. Disagreements at any of the stages were discussed and, when necessary, a third author convened to find a resolution. Some records reported results that were redundant with a later paper of the same research group, mainly because the earlier record was a conference paper or because an extended version of the research paper had been published elsewhere at a later date. When this happened, earlier and shorter reports were excluded.

Data collection process

Our original intention was to rely on the PICO framework [14] for data collection. However, given the relative youth and heterogeneity of the research field reviewed, and the lack of existing reviews on the topic, we adapted a data extraction tool specifically for our purposes. This tool took the form of four comprehensive tables which were used to extract the relevant information from each paper. Those tables summarise general study information, data details, methodology and clinical applicability.

The tables were initially "piloted" with a few studies, in order to ensure they were fit to purpose. Information extraction was performed independently by two reviewers and consistency was compared. When differences about extracted items was not resolved by discussion, the third author was available to mediate with the paper's full text as reference.

Data items (extraction tool)

As stated in the data collection process, data items will be extracted through the elaboration of four tables. These tables are:

- **SPICMO:** inspired in the PICO framework, it contains information on Study, Population, Interventions, Comparison groups, Methodology and Outcomes. More details can be found just before Table 5 (supplementary material).
- **Data details:** dataset/subset size, data type, other data modalities, data annotation, data availability and language. More details can be found just before Table 6 (supplementary material).
- **Methodology details:** pre-processing, features generated, ML task/method, evaluation technique and results. More details can be found just before Table 7 (supplementary material).
- **Clinical applicability:** research implications, clinical potential, risk of bias, and strengths/limitations. More details can be found just before Table 8 (supplementary material).

Risk of bias in individual studies

Many issues, such as bias, do not apply straightforwardly to this review because it focuses on diagnostic and prognostic test accuracy, rather than interventions. Therefore, if there were to be significance tests they would be for comparisons between the results of the different methods. Besides, the scope of the review is machine learning technology, where the evaluation through significance testing is rare. Papers that rely exclusively on traditional statistics will be excluded, and therefore we expect the review to suffer from a negligible risk of bias in terms of significance testing.

The risk of bias in machine learning studies often comes from how the data is prepared in order to train your models. In a brief example, if a dataset is not split in a training and a testing subset, the model will be trained and tested on the same data. Such model is likely to achieve very good results, but chances are that its performance will drop dramatically when tested on unseen data. This risk is called "overfitting", and is assessed in the last table of the review (Clinical Significance, Table 8). Other risks accounted for in this table are data balance, the use of suitable metrics, the contextualization of results and the sample size. Data balance reports whether the dataset has comparable numbers of AD and healthy participants, as well as in terms of gender or age. Suitable metrics is an assessment of whether the metric chosen to evaluate a model is appropriate, in conjunction with data balance and sample size (e.g. accuracy is not a robust metric when a dataset is imbalanced). Contextualization refers to whether their study results are compared to a suitable baseline (i.e. a measure without a target variable or comparable research results). Finally, sample size is particularly relevant because machine learning methodology was developed for large datasets, but data scarcity is a distinctive feature of this field.

The poor reporting of results and subsequent interpretation difficulties is a longstanding challenge of diagnostic test accuracy research [15]. Initially, we considered two tools for risk of bias assessment, namely the "QUADAS-2: Quality Assessment of Diagnosis Studies checklist - 2" [16] and the "PROBAST: Predic-

tion model Risk Of Bias ASsessment Tool" [17]. However, our search covers an emerging interdisciplinary field where papers are neither diagnostic studies nor predictive ones. Additionally, the Cochrane Collaboration recently emphasised a preference for systematic reviews to focus on the performance of individual papers' on the different risk of bias criteria [18]. Consequently, we decided to assess risk of bias as part of the Clinical Applicability table (table 8), according to criteria that are suitable to the heterogeneity currently inherent to the field. These criteria include the risks of bias described above, as well as an assessment of generalisability, replicability and validity, which are standard indicators of the quality of a study. Risk of bias was independently assessed by two reviewers and disagreements were resolved by discussion.

Data synthesis

Given the discussed characteristics of the field, as well as the broad range of details covered by the tables, we anticipate a thorough discussion of all the deficiencies and inconsistencies that future research should address. Therefore, we summarise the data in narrative form, following the structure provided by the features summarised in each table. Although a meta-analysis is beyond scope at the current stage of the field, we do report outcome measures in a comparative manner when possible.

Confidence in cumulative evidence

We will assess accuracy of prognostic and diagnostic tools, rather than confidence in an intervention. Hence, we will not be drawing any conclusions related to treatment implementation.

Background on AI, Cognitive tests and Databases

This section briefly defines key terminology and abbreviations referring and offers a taxonomy of features, adapted from Voleti et al. [19], to enhance the readability of the systematic review tables. This section also briefly describes the most commonly used databases and neuropsychological assessments, with the intention of making these accessible for the reader.

AI, machine learning, and speech technologies

AI can be loosely defined as a field of research that studies artificial computational systems that are capable of exhibiting human-like abilities or human level performance in complex tasks. While the field encompasses a variety of symbol manipulation systems and manual encoding of expert knowledge, the majority of methods and techniques employed by the studies reviewed here concern machine learning methods. While machine learning dates back to the 50's, the term "machine learning" as it is used today, originated within the AI community in the late 70's to designate a number of techniques designed to automate the process of knowledge acquisition. Theoretical developments in computational learning theory and the resurgence of connectionism in the 80's helped consolidate the field, which incorporated elements of signal processing, information theory, statistics and probabilistic inference, as well as inspiration from a number of disciplines.

The general architecture of a machine learning system, as used in AD prediction based on speech and language can be described in terms of the learning task, data representation, learning algorithm, nature of the "training data" and performance measures. The learning task concerns the specification of the function to be learnt by the system. In this review, such functions include classification (for instance, the mapping of a voice or textual sample from a patient to a target category such as "probable AD", "MCI" or "healthy control") and regression tasks (such as mapping the same kind of input to a numerical score, such as a neuropsychological test score). The data representation defines which features of the vocal or linguistic input will be used in the mapping of that input to the target category or value, and how these features will be formally encoded. Much research in machine learning applied to this and other areas focuses on data representation. A taxonomy of features used in the papers reviewed here is presented on table 1. There is a large variety of learning algorithms available to the practitioner, and a number of them have been employed in AD research. These range from connectionist systems, of which most "deep learning" architectures are examples, to relatively simple linear classifiers such as naïve Bayes and logistic regression, to algorithms that produce interpretable outputs in the form of decision trees or logical expressions, to ensembles of classifiers and boosting methods. The nature of the training data affects both its representation and the choice of algorithm. Usually, in AD research, patient data are annotated with labels for the target category (e.g. "AD", "control") or numerical scores. Machine learning algorithms that make use of such annotated data for induction of models are said to perform supervised learning, while learning that seeks to structure

unannotated data is called unsupervised learning. Performance measures, and by extension the loss function with respect to which the learning algorithm attempts to optimise, usually depend on the application. Commonly used performance measures are accuracy, sensitivity (also known as recall), specificity, positive predictive value (also known as precision), and summary measures of trade-offs between these measures, such as area under the receiver operating characteristic curve, and F scores. These methods and metrics are further detailed below.

Cognitive tests

This is a brief description of the traditional cognitive tests (as opposed to speech-based cognitive tasks) most commonly applied in this field, with two main purposes. On the one hand, neuropsychological assessments are one of the several factors on which clinicians rely in order to make a clinical diagnosis, which in turn results on participants being assigned to an experimental group (i.e. healthy control, SCI, MCI, or AD). On the other hand, some of these tests are recurrently used as part of the speech elicitation protocols.

Batteries used for diagnostic purposes consist of reliable and systematically validated assessment tools that evaluate a range of cognitive abilities. They are specifically designed for dementia and aimed to be time-efficient, as well as able to highlight preserved and impaired abilities. The most commonly used batteries are the Mini-Mental State Examination [MMSE; 26], the Montreal Cognitive Assessment [MoCA; 27], the Hierarchical Dementia Scale-Revised [HDS-R; 28], the Clinical Dementia Rating [CDR; 29], the Clock Drawing Test [CDT; 30], the Alzheimer's Disease Assessment Scale, Cognitive part [ADAS-Cog 31], the Protocol for an Optimal Neurpsychological Evaluation [32, PENO, in French;] or the General Practitioner Assessment of Cognition [GPCog; 33]. Most of these tests have been translated into different languages, such as the Spanish version of the MMSE [MEC; 34], which is used in a few reviewed papers.

Tools measuring general functioning, such as the General Deterioration Scale [GDS; 35] or Activities of Daily Living, such as the Katz Index [36] and the Lawton Scale [37], are also commonly used. Based on the results of these tests, clinicians usually proceed to diagnose MCI, following Petersen's criteria [38], or AD, following NINCDS-ADRDA criteria [39]. Alternative diagnoses appear in some texts, such as Functional Memory Disorder (FMD), following Schmidtke et al. [40]'s criteria.

Speech elicitation protocols often include tasks extracted from examinations that were originally designed for aphasia, such as fluency tasks. Semantic verbal fluency tasks [SVF, in COWAT; 41] and are often known as "animal naming" because they require the participant generating a list of nouns from a certain category (e.g. animals) while being recorded. Another tool recycled from aphasia examinations is the Cookie Theft Picture task [42], which requires participants to describe a picture depicting a dynamic scene, and hence to also elaborate a short story. Although that is by far the most common picture used in such tests, other pictures have also been designed to elicit speech in a similar way [e.g. 43].

Another group of tests consists, essentially, of language sub-tests (i.e. vocabulary) and immediate/delayed recall tests, extracted from batteries to measure intelligence and cognitive abilities, such as the Wechsler Adult Intelligence Scale [WAIS-III; 44] or the Wechsler Memory Scale [WMS-III; 45], respectively. Besides, the National Adult Reading Test [NART; 46], the Arizona Battery for Communication Disorders of Dementia ABCD battery [ABCD; 47], the Grandfather Passage, [48] and a passage of The Little Prince [49] are also used to elicit speech in some articles.

Databases

Although types of data will be further discussed later, we hereby give an overview of the main datasets described. For space reasons, we only mention here those datasets which have been used in more than one study, and for which a requesting procedure might be available. For monologue data:

- *Pitt Corpus:* by far the most commonly used. It consists of picture descriptions elicited by the Cookie Theft Picture, generated by healthy participants and patients with probable AD, and linked to their neuropsychological data (i.e. MMSE). It was collected by the University of Pittsburgh [50] and distributed through DementiaBank [51].
- *BEA Hungarian Dataset:* this is a phonetic database, containing over 250 hours of multipurpose Hungarian spontaneous speech. It was collected by the Research Institute for Linguistics at the Hungarian Academy of Sciences [52] and distributed through META-SHARE.
- *Gothenburgh MCI database:* this includes comprehensive assessments of young elderly participants during their Memory Clinic appointments and senior citizens that were recruited as their healthy counterparts [53]. Speech research un-

Table 1
Feature taxonomy, adapted from Voleti et al. [19].

Category	Subcategory	Feature type	Feature name, abbreviation, reference.
Text-based (NLP)	Lexical features	Bag of words, vocabulary analysis	*BoW*, *Vocab*.
		Linguistic Inquiry and Word Count	*LIWC* [20]
		Lexical diversity	Type-Token Ratio (*TTR*), Moving Average TTR (*MATTR*), Simpson's Diversity Index (*SDI*) Brunét's Index (*BI*), Honoré's Statistic (*HS*).
		Lexical Density	Content density (*CD*), Idea Density (*ID*), *P*-Density (*PD*).
		Part-of-Speech tagging	*PoS*.
	Syntactical features	Constituency-based parse tree scores	*Yngve* [21], *Frazier* [22].
		Dependency-based parse tree scores	
		Speech graph	Speech Graph Attributes (*SGA*).
	Semantic features	Matrix decomposition methods	Latent Semantic Analysis (*LSA*), Principal Component Analysys (PCA).
	(Word and sentence embeddings)	Neural word/sentence embeddings	*word2vec* [23]
		Topic modelling	*Latent Dirichlet Allocation* [24].
		Psycholinguistics	Reliance on familiar words (*PsyLing*).
	Pragmatics	Sentiment analysis	*Sent*.
		Use of language *UoL*	Pronouns, paraphrasing, filler words (*FW)*.
		Coherence	*Coh*.
Acoustic	Prosodic features	Temporal	Pause rate (*PR*), Phonation rate (*PhR*), Speech rate (*SR*), Articulation rate (*AR*). Vocalisation events.
		Fundamental Frequency	F_0 and trajectory.
		Loudness and energy	*loud*, *E*.
		Emotional content	*emo*.
	Spectral features	Formant trajectories	F_1, F_2, F_3.
		Mel Frequency Cepstral Coefficients	*MFCCs* [25].
	Vocal quality	Jitter, Shimmer, harmonic-to-noise ratio	*jitt*, *shimm*, *HNR*.
	ASR-related	Filled pauses, repetitions, dysfluencies, hesitations. fractal dimension, entropy.	*FP, rep, dys, hes, FD, entr*.
		Dialogue features (i.e. Turn-Taking)	*TT*:avg turn length, inter-turn silences.

dertaken with this dataset uses the Cookie Theft picture description and reading tasks subsets, all recorded in Swedish.

For dialogue data, the *Carolina Conversations Collection (CCC)* is the only available database. It consists of conversations between healthcare professionals and patients suffering from a chronic disease, including AD. For dementia research, participants are assigned to an AD group or a non-AD group, if their chronic condition is unrelated to dementia (i.e. diabetes, heart disease). Conversations are prompted by questions about their health condition and experience in healthcare. It is collected and distributed by the Medical University of South Carolina [54].

In addition, some of the reviewed articles refer to the *IVA dataset*, which consists of structured interviews undertaken and recorded simultaneously by an Intelligent Virtual Agent (a computer "avatar") [55]. However, the potential availability of this dataset is unknown.

Results

Adding up all digital databases, the searches resulted in 3,605 records. Another 43 papers were identified by searching through bibliographies and citations and 6 through research portal suggestions, adding up to 3,654 papers in total. Of those, 306 duplicates were removed using EndNote X8, leaving 3,348 for the first screening phase. In this first phase, 3,128 papers were excluded based on title and abstract, and therefore 220 reached the second phased of screening. Five of these papers did not have a full-text available, and therefore 215 papers where fully screened. Finally, 51 papers were included in the review (Figure 2).

Existing literature

The review by Voleti et al. [19] is to our knowledge, the only published work with a comparable aim to the the present review, although there are important scope differences. First of all, the review by Voleti et Al. differs from ours in terms of methodological scopes. Whilst their focus was to create a taxonomy for speech and language features, ours was to survey diagnosis and cognitive assessment methods that are used in this field and to assess the extent to which they are successful. In this sense, our search was intentionally broad. There are also differences in the scope of medical applications. Their review studies a much broader range of disorders, from schizophrenia to depression and cognitive decline. Our search, however, targeted cognitive decline in the context of dementia and Alzheimer's Disease. It is our belief that these reviews complement each other in providing systematic accounts of these emerging fields.

Data extraction

Tables with information extracted from the papers are available as supplementary material. There are 4 different tables: a general table concerning usual clinical features of interest (after the PICOS framework), and three more specific tables concerning data details, methodology details and implications for clinicians and researchers. Certain conventions and acronyms were adopted when extracting article information, and should be considered when interpreting the information contained on those tables. These conventions are available in the supplementary material, prior to the tables.

Discussion

In this section, the data and outcomes of the different tables are synthesized in different subsections and put into perspective. Consistent patterns and exceptions are outlined. Descriptive aspects are organised by column names, following table order and referencing their corresponding table in brackets.

Study aim and design (table 5: SPICMO)

Most of the reviewed articles aim to use acoustic and/or linguistic features in order to distinguish the speech produced by healthy participants from the one produced by participants with a certain degree of cognitive impairment. The majority of studies attempt binary models to detecting AD and, less often, MCI, in comparison to HC. A few studies also attempt to distinguish between MCI an AD. Even when the dataset contains three or four groups (e.g. HC, SCI, MCI, AD), most studies only report pairwise group comparisons [57–61]. Out of 51 reviewed papers, only seven did attempt three-way [49, 62–64] or four-way [12, 65, 66] classification. Their results are inconclusive and present potential biases related to the quality of the datasets (i.e. low accuracy on balanced datasets, or high accuracy on imbalanced datasets).

Slightly different objectives are described by Clark et al. [67], the only study predicting conversion from MCI to AD, and by Weiner and Schultz [68], the only study predicting progression from HC to any form of cognitive impairment. While these studies also learnt classifiers to detect differences between groups, they differ from other studies in that they use longitudinal data. There is only one article with a different aim than classification. This is the study by Duong et al. [69], who attempt to describe AD and HC discourse patterns through cluster analysis.

Despite many titles mentioning cognitive monitoring, most research addresses only the presence or absence of cognitive impairments (41, out of 51 papers). Outside of those, seven papers are concerned with three or four disease stages [12, 49, 62–66], two explore longitudinal cognitive changes (although still through binary classification) [67, 68] and one describes discourse patterns [69]. We note that future research could take further advantage of this longitudinal aspect to build models able to generate a score reflecting risk of developing an impairment.

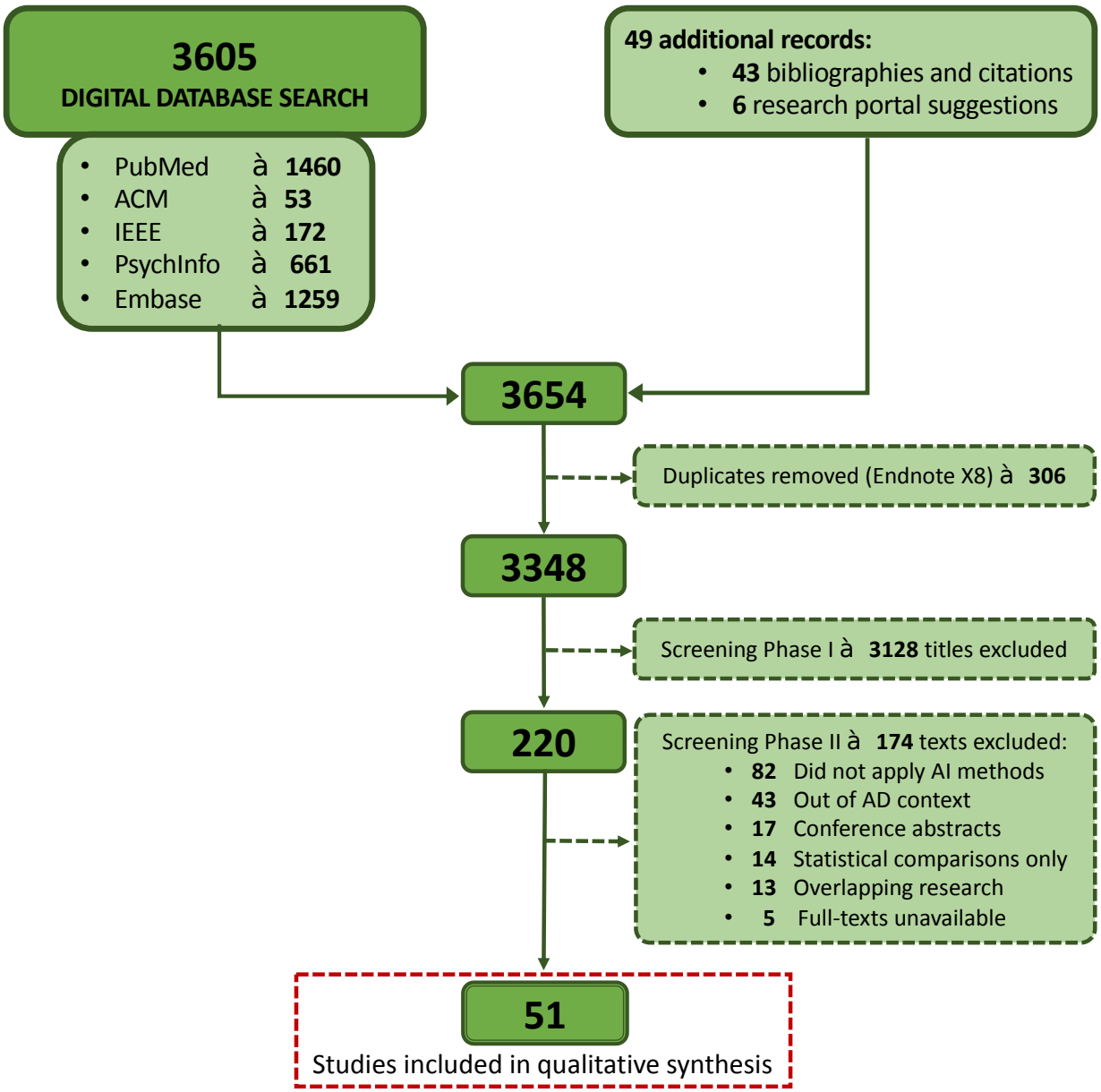

Fig. 2. Screening and selection procedure, following guidelines provided by PRISMA [56].

Population (table 5: SPICMO)

The target population are elderly people who are healthy or exhibit certain signs of cognitive decline related to AD (i.e. SCI, MCI, AD). Demographic information is frequently reported, most commonly age, followed by gender and years of education.

Cognitive scores such as MMSE are often part of the descriptive information provided for study participants as well. This serves group assignment purposes and allows quantitative comparisons of participants' degree of cognitive decline. In certain studies, MMSE is used to calculate the baseline against which classifier performance will be measured [43, 70, 71]. However, despite being widely used in clinical and epidemiological investigations, MMSE has been criticised for having ceiling effects, especially when used to assess preclinical AD [72].

Some studies report no demographics [6, 66, 68, 73, 74], only age [10, 75], only age and gender [61, 76,

77], or only age and education [70, 78]. An exception is the dataset AZTIAHORE [79, 80], which contains the youngest healthy group (20-90 years old) and a typical AD group (68-98 years old), introducing potential biases due to this imbalance. Demographic variables are established risk factors for AD [81], therefore demographics reporting is essential for this type of study.

Interventions (table 5: SPICMO)

Study interventions almost invariably consist of a speech generation task preceded by a health assessment. This varies between general clinical assessments, including medical and neurological examinations, and specific cognitive testing. The comparison groups are based on diagnosis groups, which in turn are established with the results of such assessments. Therefore, papers lacking that information do not specify their criteria for group assignment [60, 61, 75, 79, 80, 82–85]. This could be problematic, since the field currently revolves around diagnostic categories, trying to identify such categories through speech data. Consequently, one should ensure that standard criteria have been used and that models are accurately tuned to these categories.

Speech tasks are sometimes part of the health assessment. For instance, speech data are often recorded during the language sub-test of a neuropsychological battery (e.g. verbal fluency, story recall or picture description tasks). Another example of speech generated within clinical assessment is the recording of patient-doctor consultations [8, 85, 86] of cognitive examinations [e.g. MMSE, 83]. There are also studies where participants are required to perform language tests outwith the health assessment, for speech elicitation purposes only. Exceptionally, two of these studies work with written rather than spoken language [87, 88]. Alternative tasks for this purpose are reading text passages aloud [e.g. 89], recalling short films [e.g. 63], retelling a story [e.g. 90], retelling a day or a dream [e.g. 91], or taking part in a semi-standardised [e.g. 68] or conversational [e.g. 10] interview.

Most of these are examples of constrained, laboratory-based interventions, which seldom include spontaneously generated language. There are advantages to collecting speech under these conditions, such as ease of standardisation, better control over potential confunding factors, and focus on high cognitive load tasks that may be more likely to elicit cognitive deficits. However, analysis of spontaneous speech production and natural conversations also has advantages. Spontaneous and conversational data can be captured in natural settings over time, thus mitigating problems that might affect performance in controlled, cross-sectional data, such as a participant having an "off day" or having slept poorly the night before the test.

Comparison groups (table 5: SPICMO)

This review targets cognitive decline in the context of AD. For its purpose, nomenclature heterogeneity has been homogenised into four consistent groups: HC, SCI, MCI and AD; with an additional group, CI, to account for unspecified impairment (see Table 2). As an exception to this nomenclature are Mirheidari et al. [8, 12, 86], who compare participants with an impairment caused by neurodegenerative disease (ND group, including AD) to an impairment caused by functional memory disoders (FMD); and Weiner and Schultz [68], Weiner et al. [74], who introduce a category called age-associated cognitive decline (AACD).

Furthermore, some studies add subdivisions to these categories. For instance, there are two studies that classify different stages within the AD group [79, 80]. Another study divides the MCI group between amnesic single domain (aMCI) and amnesic multiple domain (a+mdMCI), although classification results for two groups are not very promising [57]. Within-subject comparisons have also been attempted, comparing participants who remained in a certain cognitive status to those who changed [67, 74].

Most studies target populations where a cohort has already been diagnosed with AD or a related condition, looking for speech differences between those and healthy cohorts. Therefore, little insight is offered into pre-clinical stages of the disease.

Outcomes of interest (table 5: SPICMO)

Given the variety of diagnostic categories and types of data and features used, it is not easy to establish state-of-the-art performance. For binary classification, the most commonly attempted task, the reported performance ranges widely depending in the data use, the recording conditions, and the variables used in modelling. For instance Lopez-de Ipiña et al. [80] reported an accuracy that varied between 60% and 93.79% using only acoustic features that were generated *ad hoc*. Although the second figure is very promising, their dataset is small, 40 participants, and remarkably imbalanced in terms of both diagnostic class and age. In

terms of class, even though they initially report 20 AD and 20 HC, the AD group is divided in three different severity stages, with 4, 10 and 6 participants respectively, whereas the control group remains unchanged (20). In terms of age, 25% percent of their healthy controls fall within a 20-60 years old age range, while 100% of the AD group are over 60 years old. In contrast, Haider et al. [11] reported 78.7% accuracy, using also acoustic features only, but generated from standard feature sets that had been developed for computational paralinguistics. Besides, this figure appears as more robust because the dataset is much larger (164 participants) and it is balanced for class, age and gender, as well as audio enhanced. Guo et al. [92] obtained 85.4% accuracy on the same dataset as [11], but using text-based features only and without establishing class, age or gender balance. All the figures quoted so far refer to monologue studies. The state-of-the-art accuracy for dialogue data is 86.6%, obtained by Luz et al. [10] using acoustic features only.

Regarding other classification experiments, we see that Mirzaei et al. [49] reports 62% for a 3-way classification, discriminating HC, MCI, AD. They are also among the few to appropriately report accuracy, since they work with a class-balanced dataset, while many other studies report overall accuracy in class-imbalanced datasets. Accuracy figures can be very misleading in the presence of class imbalance. A trivial rejector (i.e. a classifier that trivially classifiers all instances as negative with respect to a class of interest), would achieve very high accuracy on a dataset that contained, say, 90% negative instances. For example, Nasrolahzadeh et al. [65] report really high accuracy with a 4-way classifier, 97.71%, but in a highly imbalanced dataset. However, Mirheidari et al. [12] reported 62% accuracy and 0.815 AUC for a 4-way classifier in a slightly more balanced dataset. Thomas et al. [66] also 4-way, only 50%, on four groups of MMSE scores. Other studies attempting 3-way classification experiments in balanced datasets are Egas López et al. [62], 56% and Gosztolya et al. [63] with 66.7%. Kato et al. [64], however, reports 85.4% 3-way accuracy in an imbalanced dataset.

These results are diverse, and it stands clear that some will lead to more robust conclusions than others. Notwithstanding, numerical outcomes are always subject to the science behind them, the quality of the datasets and the rigour of the method. This disparity of results therefore highlights the need for improved standards of reporting in this kind of study. Reported results should include metrics that allow the reader to assess the trade-off between false positives and false negatives in classification, such as specificity, sensitivity, fallout and F scores, as well measures that are less sensitive to class imbalance, widely used in other applications of computational paralinguistics, such as unweighted average recall. Contingency tables and ROC curves should also be provided whenever possible. Given the difficulties in reporting, comparing and differentiating the results for the 51 reviewed studies on an equal footing, we refer the reader to Tables 5 and 7, available in the suplementary material for details.

Size of dataset or subset (table 6: Data Details)

Within a machine learning context, all the reviewed studies use relatively small datasets. About 31% train their models with less than 50 participants [8, 10, 49, 64, 68, 71, 79, 80, 82, 83, 85, 86, 89, 91, 93, 94], whilst only 27% have 100 or more participants [9, 11, 57, 60, 67, 70, 73, 75–77, 92, 95–97]. In fact, 5 report samples with less than 30 participants [68, 79, 83, 89, 94]

It is worth noting that those figures represent the dataset size in full, which is then divided in two, three or four groups, most of the times unevenly. There are only 6 studies where not only the dataset, but also each experimental group contains 100 or more participants/speech samples [6, 9, 11, 85, 92, 95]. All of these studies used the *Pitt Corpus*.

The *Pitt Corpus* is the largest dataset available. It is used in full by Ben Ammar and Ben Ayed [95], and contains 484 speech samples, although it is not clear to how many unique participants these samples belong. With the same dataset, Luz [6] reports 398 speech samples, but again, no number of unique participants. However, another study working with the *Pitt Corpus* does report 473 speech samples from 264 participants [9]. It is important for studies to report numbers of unique participants in order to allow the reader to assess the risk that the ML models might actually be simply learning to recognise participants rather than their underlying cognitive status. This risk can be mitigated, for example, by ensuring that all samples from each participant are in either the training set or the test set, but not both.

Data type (table 6: Data Details)

This column refers to the data used in each reviewed study, indicating if these data consist of monologues or dialogues, purposefully elicited narratives or speech obtained through a cognitive test. It also in-

cludes whether data was recorded or recorded and transcribed, and how this transcription was done (i.e. manual or automatic).

Of the reviewed studies, 82% used monologue data, and most of them (36) obtained speech through a picture description task (e.g. *Pitt Corpus*). These are considered relatively spontaneous speech samples, since participants may describe the picture in whichever way they want, although the speech content is always constrained. Among other monologue studies, eight work with speech obtained through cognitive tests, frequently verbal fluency tasks. Only two papers rely on truly spontaneous and natural monologues, prompted with an open question instead of a picture description [60, 65].

Dialogue data are present less frequently, in 27% of the studies, and elicited more heterogeneously. For instance, in structured dialogues (4 studies), both speakers (i.e. patient and professional) are often recorded while taking a cognitive test [8, 12, 83, 94]. Semi-structured dialogues (5 studies) are interview-type conversations where questions are roughly even across participants. From our point of view, the most desirable data type are conversational dialogues (5 studies), where interactive speech is prompted with the least possible constraints [10, 66, 79, 80, 98]. A few studies have collected dialogue data through an intelligent virtual agent (IVA) [8, 12, 94] showing the potential for data to be collected remotely, led by an automated computer system.

In terms of data modalities (e.g. audio, text or both), two studies are the exception where data was directly collected as written text [87, 88]. A few studies (6) work with audio files and associated ASR transcriptions [12, 43, 62, 63, 77, 99]. Another group of studies (14), use solely voice recordings [49, 57, 60, 61, 64, 65, 71, 79, 80, 82, 84, 89, 97, 100]. More than half of the studies (55%) rely, at least partially, on manually transcribed data. This is positive for data sharing purposes, since manual transcriptions are usually considered golden standard quality data. However, methods that rely on transcribed speech may have limited practical applicability, as they requires costly and time-consuming, and often (as when ASR is used) error prone (see section on pre-processing below) intermediate steps compared to working directly with the audio recordings.

Other modalities (table 6: Data Details)

The most frequently encountered data modality, apart from speech and language, is structured data related to cognitive examinations, largely dominated by MMSE and verbal fluency scores. Another modality is video, which is available in some datasets such as CCC [10, 98], AZTITXIKI [79], AZTIAHORE [60, 80], IVA [12, 85] or the one in Tanaka et al. [94], although it is not included in their analysis. Other analysed modalities include neuroimaging data, such as MRI [67] and fNIRS [64], eye-tracking [7, 94] or gait information [71].

In order to develop successful prediction models for pre-clinical populations, it is likely that future interactive AI studies will begin to include demographic information, biomarker data and lifestyle risk factors [101].

Data annotation (table 6: Data Details)

Group labels and sizes are presented in this section of the Data Details table, the aim of which is to give information about the available speech datasets. Accordingly, labels remain as they are reported in each study, as opposed to the way in which we homogenised them to describe Comparison Groups in Table 5. In other words, even though the majority of studies annotate their groups as HC, SCI, MCI and AD, some do not. For example, the HC group is labelled as CON (control) [91], NC (normal cognition) [57, 64, 88, 99], CH (cognitively healthy) [82], and CN (cognitively normal) [67]. SCI can also be named SMC [96], and there is a similar but different category (AACD) reported in two other studies [68, 74]. MCI and AD are more homogeneous due to being diagnostic categories that need to meet certain clinical criteria to be assigned, although some studies do refer to AD as *dementia* [62, 95]. Another heterogeneous category is CI (i.e. unspecified cognitive impairment), which is annotated as *low* or *high* MMSE scores [93], or as *mild dementia* [89]. *Mild dementia* may sound similar to MCI, however the study did not report diagnostic criteria for MCI to be considered.

This section offers insight into another aspect in which lack of consensus and uniformity is obvious. Using accurate terminology (i.e. abiding by diagnosis categories) when referring to each of these groups could help establish the relevance of this kind of research to clinical audiences.

Data balance (table 6: Data Details)

Only 39% (20) of the reviewed studies present class balance, that is, the number of participants is evenly distributed across the two, three or four diagnostic categories [7, 8, 11, 49, 60, 62–64, 75, 78, 82, 84, 86, 88–91, 94, 95, 98]. Among these 20 studies, one reports only between-class age and gender balance [94], another one reports class balance, within-class gender balance and between-class age and gender balance [11]. A few report balance for all features except for within-class gender balance, which is not specified [62, 63, 88]. Lastly, there is only one study that, apart from class balance, also reports gender balance within and between classes, as well as age and education balance between classes [87]. Surprisingly, nine other studies fail to report one or more demographic aspects.

Sometimes gender is reported per dataset, but not per class [e.g. 93], and therefore not accounted for in the analysis, even though is one of the main risk factors [81]. Often, *p*-values are appropriately presented to indicate that demographics are balanced between groups [e.g. 62]. Unfortunately, almost as often, no statistical values are reported to argue for balance between groups [e.g. 83]. There are also cases where where the text reports demographic balance but neither group distributions nor statistical tests are presented [e.g. 91]. Another aspect to take into account is the differences between raw and pre-processed data. For instance, Lopez-de Ipiña et al. [79, 80] describe a dataset where 20% of the HC speech data, but 80% of the AD speech data, is removed during pre-processing. Hence, even if these datasets had been balanced before (they were not) they will definitely not be balanced after pre-processing has taken place.

It is also worth discussing the reasons behind participant class imbalance when the same groups are class balanced in terms of samples. Fraser et al. [9], for example, work with a subset of the *Pitt Corpus* of 97 HC participants and 176 AD participants, however, the number of samples is 233 and 240, respectively. Similar patterns apply to other studies where the number of participants and samples are reported [92, 98]. Did HC come for more visits, or did perhaps AD participants fail to come to later visits or drop out of the study? These incongruities could be hiding systematic group biases.

Conclusions drawn from imbalanced data are subject to a greater probability of bias, especially in small datasets. For example, certain performance metrics to evaluate classifiers are more robust (e.g. *F1*) than others (e.g. *acc*) against this imbalance. Accordingly, in this table, the smaller the dataset, the more strict we have been when evaluating the balance of its features. Moving forward, it is desirable that more emphasis is placed on data balance, not only in terms of group distribution, but also in terms of those demographic features established risk factors (i.e. age, gender and years of education).

Data availability (table 6: Data Details)

Strikingly, very few studies make their data available, or even report on its (un)availability, even when using available data hosted by a different institution (e.g. studies using the *Pitt Corpus*). The majority (77%, 39 studies) fail to report on data availability. From the remaining 12 studies, nine use data from DementiaBank (*Pitt Corpus* or *Mandarin_Lu*) and do report data origin and availability. However, only [75, 90] share the exact specification of the subset of *Pitt Corpus* used for their analysis, in order for other researchers to be able to replicate their findings, taking advantage of the availability of the corpus. The same applies to Luz et al. [10], who made available their identifiers for the CCC dataset. One other study, Fraser et al. [7], mentions that data are available upon request to authors.

Haider et al. [11], one of the studies working on the *Pitt Corpus*, has released their subset as part of a challenge for INTERSPEECH 2020, providing the research community with a dataset matched for age and gender and with enhanced audio. In such an emerging and heterogeneous field, shared tasks and data availability are important progression avenues.

Language (table 6: Data Details)

As expected, a number of studies (41%) were conducted through English. However, there is a fair amount of papers using data in a variety of languages, including: Italian [91], Portuguese [57, 90], Chinese and Taiwanese [82], French [49, 69, 77, 96, 102], Hungarian [62, 63, 99], Spanish [83, 89, 100], Swedish [7, 59, 87], Japanese [64, 71, 94], Turkish [84], Persian [65], Greek [61, 88], German [68, 74] or reported as multilingual [60, 79, 80].

This is essential if screening methodologies for AD are to be implemented worldwide [103]. The main caveat, however, is not the number of studies conducted in a particular language, but the fact that most of the studies conducted in languages other than En-

glish do not report on data availability. As mentioned, only Dos Santos et al. [90] and Fraser et al. [7] report their data being accessible upon request, and Chien et al. [82] works with data available from DementiaBank. For speech-based methodology aimed at AD detection, it would be a helpful practice to make these data available, so that other groups are able to increase the amount of research done in any given language.

Pre-processing (table 7: Methodology)

Pre-processing includes the steps for data preparation prior to data analysis. It is essential to determine in which shape any given data is introduced in the analysis pipeline, and therefore, the outcome of it. However, surprisingly little detail is reported in the reviewed studies.

Regarding text data, the main pre-processing procedure is transcription. Transcription may happen manually or through ASR. The Kaldi speech recognition toolkit [104], for instance, was used in several recent papers [e.g. 12, 62]. Where not specified, manual transcription is assumed. Although many ASR approaches do extract information on word content [e.g. 8, 43, 71, 85, 86, 96], some focus on temporal features, which are content-independent [e.g. 63, 82]. Some studies report their transcription unit, that is, word-level transcription [e.g. 9], phone-level transcription [e.g. 63] or utterance-level transcription [e.g. 91]. Further text pre-processing involves tokenisation [73, 82, 90, 94], lemmatization [87] and removal of stopwords and punctuation [87, 90]. Depending on the research question, dysfluencies are also removed [e.g. 87, 90], or annotated as relevant for subsequent analysis [e.g. 59].

Currently, commercial ASRs are optimised to minimize errors at word level, and therefore not ideal for generating non-verbal acoustic features. Besides, it seems that AD patients are more likely to generate ungrammatical sentences, incorrect inflections and other subtleties that are not well handled by such ASR systems. In spite of this, only a few papers, by the same research group, rely on ASR and report WER (word error rate), DER (diarisation eror rate) or WDER (word diarisation error rate) [8, 12, 85]. It is becoming increasingly obvious that off-the-shelf ASR tools are not readily prepared for dementia research, and therefore some reviewed studies developed their own custom ASR systems [43, 63].

Regarding acoustic data, pre-processing is rarely reported outside the audio files being put through an ASR. When reported, it mainly involves speech-silence segmentation with voice activity deteciton algorithms (VAD), including segment length and the acoustic criterion chosen for segmentation thresholds (i.e. intensity) [6, 11, 43, 49, 60, 61, 64, 65, 68, 76, 79, 80, 96, 102]. It should also include any audio enhancement procedures, such as volume normalisation or removal of background noise, only reported in Haider et al. [11] and Sadeghian et al. [43].

We concluded from the reviewed papers that it is not common practice for authors in this field to give a complete account of the data pre-processing procedures they followed. As these procedures are crucial to reliability and replicability of results, we recommend that further research specify these procedures more thoroughly.

Feature generation (table 7: Methodology)

Generated speech features are divided into two main groups, text-based and acoustic features, and follow the taxonomy presented in Table 1. Some studies work with multimodal feature sets, including images [94] and gait [71] measurements.

Text-based features comprise a range of NLP elements, commonly a subset consisting of lexical and syntactical indices such as type-token ratio (*TTR*), *idea density* or *Yngve* and *Frazier* indices. *TTR* is a measure of lexical complexity, calculated by taking the total number of unique words, also called lexical items (i.e. types) and dividing by the total number of words (i.e. tokens) in a given language instance [105]. *Idea density* is the number of ideas expressed in a given language instance, with 'ideas' understood as new information and adequate use of complex propositions. High early *idea density* seems to be a lower risk predictor for developing AD later in life, whereas lower idea density appears associated with brain atrophy [106]. *Yngve* [21] and *Frazier* [22] scores indicate syntactical complexity by calculating the depth of the parse tree that results from the grammatical analysis of a given language instance. Both indices have been associated with working memory [107] and showed a declining pattern in the longitudinal analysis of the written work by Iris Murdoch, a novelist who was diagnosed with AD [108].

In some studies, the research question targets a specific aspect of language, such as syntactical complexity [59], or a particular way of representing it, such as speech graph attributes [57]. Fraser et al. [9] present a more comprehensive feature set, including some

acoustic features. Similar to Fraser et al. [9], although less comprehensive, a few other studies combine text-based and acoustic features [8, 43, 71, 78, 86, 89, 91, 92, 94, 96]. However, most published research is specific to one type of data or another.

The most commonly studied acoustic features are prosodic temporal features, which are almost invariably reported, followed by ASR-related features, specifically pause patterns. There is also focus on spectral features (features of the frequency domain representation of the speech signal obtained through application of the Fourier transform), which include *MFCCs* [62]. The most comprehensive studies include spectral, ASR-related, prosodic temporal, voice quality features [8, 49, 60, 79, 84, 92, 100], as well as features derived from the Higuchi Fractal Dimension [80] or from higher order spectral analysis [65]. It is worth noting here that Tanaka et al. [94] extract F_0's coefficient of variation per utterance. The decision to not extract F_0's mean and SD was due to their association with individual differences and sex. Similarly, Gonzalez-Moreira et al. [89] report F_0 and functionals in semitones, because research argues that using semitones to express F_0 reduces gender differences [109], which is corroborated by the choice of semitones in the standardised eGeMAPS [11].

Studies using spoken dialogue recordings extract turn-taking patterns, vocalisation instances and speech rate [10, 94]. Those focusing on transcribed dialogues also extract turn-taking patterns, as well as dysfluencies [8, 12, 86]. Guinn et al. [98] work with longitudinal dialogue data but do not extract specific dialogue or longitudinal features.

With regards to feature selection, 30% of the studies do not report feature selection procedures. Amongst those that do, the majority (another 30%) report using a filter approach based on a statistical index of feature differences between classes, such as *p*-values, Cohen's *d*, *AUC* or *Pearson's* correlation. Others rely on wrapper methods [49], RFE [8, 86], filter methods based on information gain [65, 95], PCA [64], best first greedy algorithm [43], and cross-validation, seeking through the iterations for which feature type contributes more to the classification model [80].

Despite certain similarities and a few features being common to most acoustic works (i.e. prosodic temporal), there is striking heterogeneity among studies. Since they usually obtain features using *ad hoc* procedures, these studies are seldom comparable, making it difficult to ascertain the state-of-the-art in terms of performance, as pointed out before, and assess further research avenues. However, this state of affairs may be starting to change as the field matures. Haider et al. [11], for instance, chose to employ standardised feature sets (i.e. ComPare, eGeMAPS, emobase) obtained through formalised procedures [110] which are extensively documented and can be easily replicated. Furthermore, one of these feature sets, eGeMAPS, was developed specifically to target affective speech and underlying physiological processes. Utilising theoretically informed, standardised feature sets increases the reliability of a study, since the same features have been previously applied (and can continue to be applied) to other engineering tasks, always extracted in the exact same way. Likewise, we argue that creating and utilising standardised feature sets will improve this field by allowing cross-study comparisons. Additionally, we recommend that the approach to feature generation should be more consistently reported to enhance study replicability and generalisability.

ML task/method (table 7: Methodology)

Most reviewed papers employ supervised learning, except for a study that uses cluster analysis to investigate distinctive discourse patterns amongst participants [69].

As regards choice of ML methods, very few papers report the use of artificial neural networks [91, 95], recurrent neural networks [82], multi-layer perceptron [43, 67, 79, 80, 86] or convolutional neural networks [60, 85]. This is probably due to the fact that most datasets are relatively small, and these methods require large amounts of data. Rather, most studies use several conventional ML classifiers, most commonly SVM, NB, RF and *k*-NN and then compare their performance. Although these comparisons must be assessed cautiously, a clear pattern seems to emerge with SVM consistently outperforming other classifiers.

Cognitive scores, particularly MMSE, are available with many datasets, including the most commonly used, *Pitt Corpus*. However, these scores mostly remain unused except for diagnostic group assignments, or more rarely, as baseline performance [43, 70, 71], in studies that conclude that MMSE is not more informative than speech based features. All supervised learning approaches work towards classification and no regression over cognitive scores is attempted. We regard this as a gap that could be explored in future research.

It is worth noting, however, that some attempts at prediction of MMSE score have been presented in workshops and computer science conferences that

are not indexed in the larger biobliography databases. These approaches achieved some degree of success. Linz et al. [111], for instance, trained a regression model that used the SVF to predict MMSE scores and obtained a mean absolute error of 2.2. A few other works used the *Pitt Corpus* for similar purposes, such as Al-Hameed et al. [112], who extracted 811 acoustic features to build a regression model able to predict MMSE scores with an average mean absolute error of 3.1; or Pou-Prom and Rudzicz [113], who used a multiview embedding to capture different levels of cognitive impairment and achieved a mean absolute error of 3.42 in the regression task. Another publication with the *Pitt Corpus* is authored by Yancheva et al. [114], who extracted a more comprehensive feature set, including lexicosyntactic, acoustic, and semantic measures, and used them to predict MMSE scores. They trained a dynamic Bayes network that modeled the longitudinal progression observed on these features and MMSE over time, reporting a mean absolute error of 3.83. This is, actually, one of the very few works attempting a progression analysis over longitudinal data.

Evaluation techniques (table 7: Methodology)

A substantial proportion of studies (43%) do not present a baseline against which study results can be compared. Amongst the remaining papers, a few set specific results from a comparable work as their baseline [6, 65] or from their own previous work [75]. Others calculate their baseline by training a classifier with all the generated features, that is, before attempting to reduce the feature set with either selection or extraction methods [83, 95, 99], with cognitive scores only [7, 43, 70, 71] or by training a classifier with demographic scores only [63]. Some baseline classifiers are also trained with a set of speech-based features that excludes the feature targeted by the research question. Some examples are studies investigating the potential of topic model features [87], emotional features [79], fractal dimension features [80], higher order spectral features [65] or feature extracted automatically, as opposed to manually [73, 85, 96]. Some studies choose random guess or naive estimations (ZeroR) [10, 11, 66, 74, 88] as their baseline performance.

While several performance metrics are often reported, *accuracy* is the most common one. While it seems straightforward to understand a classifier's performance by knowing its *accuracy*, it is not always appropriately informed. Since *accuracy* is not robust against dataset imbalances, it is only appropriate when diagnostic groups are balanced, such as when reported in Roark et al. [78], Khodabakhsh and Demiroğlu [84]. This is especially problematic for works on imbalanced datasets where accuracy is the only metric reported [9, 12, 43, 60, 66, 77, 83, 92, 93, 100]. Clinically relevant metrics such as *AUC* and *EER* [e.g. 61, 102], which summarise the rates of false alarms and false negatives, are reported in less than half of the reviewed studies.

Cross-validation (CV) is probably the most established practice for classifier evaluation. It is reported in all papers but five, of which two are not very recent [66, 93], another two do not report CV but report using a hold-out set [82, 91], and only one reports using neither CV nor a hold-out set procedure [95]. There is a fair amount of variation within the CV procedures reported, since datasets are limited and heterogeneous. For example, leave-one-out CV is used in one third of the reviewed papers, as an attempt to mitigate the potential bias caused by using a small dataset. Several other studies choose leave-pair-out CV instead [7, 70, 73, 75, 78, 97], since it produces unbiased estimates for *AUC* and also reduces potential size bias. There is also another research group who attempted to reduce the effects of their imbalance dataset by using stratified CV [68, 74]. Lastly, no studies report hold-out set procedures, except for the two mentioned above, with training/test sets divided at 80/20% and 85/15%, respectively, and another study where the partition percentages are not detailed [97].

There is a potential reporting problem in that many studies do not clearly indicate whether their models' hyper-parameters were optimized on the test set within or outside each fold of the CV. However, CV is generally considered the best method of evaluation when working with small datasets, where held-out set procedures would be even less reliable, since they would involve testing the system on only a few samples. CV is therefore an appropriate choice for the articles reviewed. The lack of systematic model validation on entirely separate datasets, and the poor practice of using accuracy as the single metric in imbalanced datasets, could compromise the generalisability of results in this field. While it is worth noting that the former issue is due to data scarcity, and therefore more difficult to address, a more appropriate selection of performance metrics could be implemented straight away to enhance the robustness of current findings.

Results overview (table 7: Methodology)

Performance varies depending on the metric chosen, the type of data and the classification algorithm used. Hence, it is very difficult to summarise these results. The evaluated classifiers range between 50% or even lower in some cases, up to over 90% accuracy. However, as we have pointed out, performance figures must be interpreted with caution due to the potential biases introduced by dataset size, dataset imbalances and non standardised *ad hoc* feature generation. . Since these biases cannot be fully accounted for and models are hardly comparable to one another, we do not think it is meaningful to further highlight the best performing models. Such comparisons will become more meaningful when all conditions for evaluation can be aligned, such as in the ADReSS challenge [115], which provides a benchmark dataset (balanced and enhanced) and commits to a reliable study comparison.

Further research on the methodology and how different algorithms behave with certain types of data will shed light on why some classifiers perform even worse than random while others are close to perfect. This could simply be because the high performing algorithms were coincidentally tested on 'easy' data (e.g. better quality, simpler structures, very clear diagnoses), but the problem could also classifier specific and therefore differences would be associated with the choice of algorithm. Understanding this would influence the future viability of this sort of technology.

Research implications (table 8: Clinical applicability)

This section reviews the papers in terms of novelty, replicability and generalisability, three aspects key to future research.

As regards **novelty**, the newest aspect of each research paper is succinctly presented in the tables. This is often conveyed by the title of an article, although caution must be exercised with regards to how this information is presented. For example, Tröger et al.'s title (2018) reads "Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment", but only when you read the full text does it become clear that such telephone screening has been simulated.

There is often novelty in pilot studies, especially those presenting preliminary results for a new project, hence involving brand new data [80, 91] or tests for a newly developed system or device [60]. Outside of those, assessing novelty in a systematic review over a 20-year span can be complicated — what was novel 10 years ago might not be novel today. For example, 3-way classification entailed novelty in Bertola et al. (2014) [57], as well as 4-way classification did in Thomas et al. (2005) [66] with text data and little success, and later in Nasrolahzadeh et al. (2018) [65] with acoustic data and an improved performance. Given its low frequency and its naturalness, we have chosen to present the use of dialogue data [10, 68, 84, 85, 94, 98] as a novelty relevant for future research. Other examples of novelty consist of automated neuropsychological scoring, either by automating traditional scoring [57] or by generating a new type of score [67, 70].

Methodological novelty is also present. Even though most studies apply standard machine learning classifiers to distinguish between experimental groups, two approaches do stand out: Duong et al.'s (2005) unique use of cluster analysis (a form of unsupervised learning) with some success, and the use of ensemble [67, 90] and cascaded [7] classifiers, with much better results. Some studies present relevant novelty for pre-processing, generating their own custom ASR systems [43, 61, 63, 99], which offers relevant insight about off-the-shelf ASR. While this is based on word accuracy, some of the customized ASR systems are phone-based [63, 99] and seem to work better with speech generated by participants with AD. Another pre-processing novelty is the use of dynamic threshold for pause behaviour [76], which could be essential for personalised screening. With regards to feature generation, "active data representation" is a novel method utilized in conjunction with standardised feature sets by Haider et al. [11], who confirmed the feasibility of a useful tool that is open software and readily available (i.e. ComParE, emobase and eGeMAPS). A particularity of certain papers is their focus on emotional response, analysed from the speech signal [79, 80]. This could be an avenue for future research, since there are other works presenting interesting findings on emotional prosody and AD [116, 117]. Last, but not least, despite the mentioned importance of early detection, most papers do not target early diagnosis, or do it in conjunction with severe AD only (i.e. if the dataset contains participants at different stages). Consequently, Lundholm Fors et al. (2018) [59] introduced a crucial novelty by not only assessing, but actively recruiting and focusing on participants at the pre-clinical stage of the disease (SCI).

Another essential novelty is related to longitudinal aspects of data [68, 77, 97]. The vast majority of studies work on monologue cross-sectional data, although some datasets do include longitudinal information (i.e. each participant has produced several speech samples). This is sometimes discarded, either by treating each sample as a different participant, which generates subject dependence across samples [74]; or by cross-observation averaging, which misses longitudinal information but does not generate this dependence [75, 97]. Other studies successfully used this information to predict change of cognitive status within-subject [68, 118]. Guinn et al. [98] work with longitudinal dialogue data that becomes cross-sectional after pre-processing (i.e. they conglomerate samples by the same participant) and they do not extract specific dialogue features.

The novelty with most clinical potential is, in our view, the inclusion of different types of data, since something as complex as AD is likely to require a comprehensive model for successful screening. However, only a few studies combine different sources of data, such as MRI data [67], eye-tracking [7], and gait [71]. Similarly, papers where human-robot interaction [8, 12, 85, 94] or telephone-based systems [96, 97] are implemented also offer novel insight and avenues for future research. These approaches offer a picture of what automatic, cost-effective screening could look like in a perhaps not so distant future.

On a different front, **replicability** is assessed based on whether the authors report complete and accurate procedures of their research. Replicability has research implications because, before translating any method into clinical practice, its performance needs to be confirmed by other researchers being able to reproduce similar results. In this review, replicabilility is labelled as *low*, *partial* and *full*. When we labelled an article as *full* with regards to replicability, we meant that their methods section was considered to be thorough enough to be reproduced by an independent researcher, from the specification of participants demographics and group size to the description of pre-processing, feature generation, classification and evaluation procedures. Only three articles were labeled as *low* replicability [66, 74, 85], as they lacked detail in at least two of those sections (frequently data information and feature generation procedures). Twenty-two and twenty-five studies were labelled as *partial* and *full*, respectively. The elements most commonly missing in the *partial* papers are pre-processing [e.g. 83] and feature generation procedures [e.g. 78], which are essential steps in shaping the input to the machine learning classifiers. It must be highlighted that all *low* replicability papers are conference proceedings, where text space is particularly restricted. Hence, it does not stand out as one of the key problems of the field, even though it is clear that the description of pre-processing and feature generation must be improved.

The last research implication is **generalisability**, which is the degree to which a research approach may be attempted with different data, different settings or real practice. Since generalisability is essentially about how translatable research is, most aspects in this last table are actually related to it:

- Whether *external validation* has been attempted is directly linked to generalisability;
- *feature balance:* results obtained in imbalanced datasets are less reliable and therefore less generalisable to other datasets;
- *contextualization of results:* for something to be generalisable is essential to know where it comes from and how does it compare to similar research;
- *spontaneous speech*: speech spontaneity is one aspect of naturalness, and the more natural the speech data, the more representative of "real" speech and the more generalisable;
- *conversational speech*: we propose that conversational speech is more representative of "real" speech;
- *content-independence:* if the classifier input includes features that are tied up with task content (e.g. lexical, syntactic, semantic, pragmatics), some degree of generalisability is lost;
- *Transcription-free:* a model that needs no transcription is free from ASR or manual transcription constraints, relying only on acoustic features. We suggest this to increase generalisability, for example, by being language-independent, therefore facilitating method usability with non-English language for which corpus training is less feasible due to even more severe data scarcity. Transcription-free methods also facilitate the protection of users' privacy, as they do not focus on speech content, which could encourage ethics committees to reduce restrictions on data collection, thereby addressing data scarcity.

Just as replicability, it is labelled as *low*, *moderate* and *high*, depending of how many of the aforementioned criteria each study meets. Different to what we described with regards to replicability, the majority of studies (20) are labelled with *low* generalisabil-

ity, 17 as *moderate* and 14 as *high*. The most common reasons for decreased generalisability are dependence on content, followed by dependence on ASR or other transcription methods, although the two are related. Content-dependence makes it difficult to apply to other tasks or data [e.g. 12, 91]. This is even more pronounced in those studies where the approach heavily relies on word content, such as *n*-grams [e.g. 75]. Linguistic models that target only one linguistic aspect are also *low* generalisability, particularly if this aspect is language-dependent [e.g. syntaxis 59]. Examples of *high* generalisability include models relying solely on acoustic features, therefore free of content and transcription constrains [e.g. 89], and especially if a standardised available feature set is used [11]. Other generalisable studies present more than one dataset [e.g. 62], different languages in the same study [e.g. 87], conversational data [e.g. 10], a system designed for direct real application [e.g. 94] and/or data from real scenarios [8].

Clinical potential (table 8: Clinical applicability)

The clinical applicability table aims to directly assess whether reviewed research could be translatable into clinical practice. Generalisability (discussed above) is essential for this purpose, but it will not be included here to avoid redundancy. We also note that clinical applicability of a diagnostic test is a somewhat vague construct in that one might need applicability in a clinical population *or* applicability for a clinician to understand its use. From a clinician's perspective, the translational steps from research on speech and language biomarkers to clinical use are not unlike those of any other diagnostic tool. This highlights, as we point out in the conclusion, that this translational development pathway would benefit from joint development between clinicians, speech and language experts, and AI experts working in concert. The other systematic aspects chosen to evaluate clinical potential are:

- **External validation:** in the majority of studies, data are collected detached from clinical practice and later analysed for result reporting. The majority of papers (84%) present neither external validation procedures nor a system design that involves them. Only four studies, all of them by the same group [8, 12, 55, 85], collect their data in a real life setting (doctor-patient consultations). Another four studies take into account feasibility for clinical screening within their system design, for example, collecting data directly with a computerised decision support system [60], through human-robot interaction [94], a computer-supported screening tool [77], or simulating telephone-based data [96].
- **Potential application:** 78% of the reviewed papers present a method that could be applied as a diagnosis support system for MCI [e.g. 87] or AD [e.g. 9]. The remaining studies work on disease progression by including SCI participants [59], predicting within-subject change [68] or discriminating between HC, MCI and AD stages [e.g. 102].
- **Global Health:** although this could include a broad range of aspects, for the purpose of this review we have chosen to mention the language of the study and the processing unit of choice. This is because most research is done in English (41%), and work published in other languages helps towards methods being more universally applicable. Also, because smaller the processing units (i.e. phoneme vs. word), tend to be more generalisable across languages. The most common processing unit is the sentence (63%), followed by conversations (16%), words (8%), syllables (4%) and phonemes (4%).
- **Remote application:** for such a prevalent disease, remote screening could significantly reduce the load on health systems. The majority of the studies, 67%, do not mention the possibility of their method being used remotely or having being designed for remote use, and only 25% suggest this as a possibility when motivating their project or discussing the results. Only four studies (2%), actually bring this into practice by experimenting with multi-modal human-robot interaction [94], infrastructure-free [77] or telephone-based [96, 97] approaches.

A further aspect, not explicitly included on this table is *model interpretability*. While the accepted opinion is that the clinicians' ability to be able to interpret an AI model is essential for the adoption of AI technologies in medicine, the issue is still the subject of lively debate, with influential machine learning researchers like Geoff Hinton arguing that "clinicians and regulators should not insist on explainability" [119]. In terms of biomarkers of disease, very few if any clinicians understand the fine detail of an MRI report; it is the results presented to them that clinicians contextualise rather than the statistical or AI journey these results

have been on to be presented to them. It could be argued that the case of speech and language biomarkers is no different. Of the papers reviewed here, only 4 mention interpretability or model interpretation explicitly [7, 9, 57, 97]. However, inherently interpretable models are used in a number of studies. Such interpretable methods were indicated in the above section on AI methods and include: linear regression, logistic regression, generalised linear and additive models, decision trees, decision rules, RuleFit, naive Bayes and K-Nearest neighbors [120], and in some cases linear discriminant analysis. As shown on Table 7, 57% of the studies reviewed included at least one of these types of models in their evaluation, even though most such inclusions were made for comparison purposes.

As regards the selected criteria, the result tables highlight that research undertaken using non-English speech data almost invariably includes acoustic features, either as part of a larger feature set, such as Beltrami et al. [91] in Italian; or exclusively relying on acoustic features, such as Nasrolahzadeh et al. [65] in Persian, Weiner and Schultz [68] in German, Lopez-de Ipiña et al. [80], Espinoza-Cuadros et al. [83], Gonzalez-Moreira et al. [89], Meilan et al. [100] in Spanish and [64] in Japanese. Apart from English, only Portuguese and Chinese have been researched exclusively with text features [57, 82, 90].

Some of the field's needs clearly arise here. Firstly, there is a need for actual attempts to use these models in real clinical practice. For twenty years, conclusions and future directions of these research papers have suggested this, but very few published studies do bring it into a realisation. Secondly, there is a need for enhanced focus on disease progression and risk prediction. Most studies mention the need for AD to be diagnosed earlier than it is now, and yet not many do actually work in that direction. Thirdly, further investment on research performed on languages other than English is needed, and increased focus on smaller language units, which are more generalisable to other languages or other samples of the same language. Alternatively, we suggest that a shift towards acoustic features only would potentially address these difficulties. Finally, one of the most obvious advantages of using Artificial Intelligence for cognitive assessment is the possibility of using less infrastructure and less personnel. In order for this to become a reality, the remote applicability of these methods requires more extensive research.

Risk of bias (table 8: Clinical applicability)

This column highlights sources of potentially systematic errors or other circumstances that may introduce bias in the inferred conclusions. These can be summarised as follows:

- **Feature balance:** class, age, gender and education balances are essential for experimental results to be unbiased. Only 13 studies (25%) are balanced for these main features, and another five are balanced in terms of class but not in terms of other features. In the studies that seek to address class imbalance in their datasets, the main strategies used are subsampling [11, 60], use of statistical methods such as stratified CV [74], and careful choice of evaluation methods including use of the UAR metric [63] and ROC curve analysis [97].
- **Suitable metrics:** equally important for bias prevention is choosing the right performance metrics to evaluate machine learning classifiers. For example, with a class-imbalanced dataset, accuracy is not a robust metric and should therefore not be used, or at least, complemented with other measures. However, 18 studies (35%) working with an imbalanced dataset report accuracy only.
- **Contextualized results:** referring mainly to whether the reported research is directly and quantitatively compared to related works, or, ideally, whether a baseline against which results can be compared is provided. Only 61% of the studies reviewed provide such context.
- **Overfitting:** studies would apply both CV and held-out sets to ensure their models do not overfit. CV should be applied when tuning ML hyperparameters when training the model, and the held-out set should be used to test the model on strictly unseen data. The majority of the studies do report CV (78%), but even more studies (90%) do not report hold-out set. Hence, there is a high risk that the reviewed models are, to some degree, overfitted to the data they have been trained with. Ideally, models should also be validated on entirely separate datasets. Only one of the studies reviewed carries out this kind of validation, although their method aims to use speech alignment in order to automatically score a cognitive task, instead of investigating the potential for dementia prediction of the linguistic or speech features themselves [70].

– **Sample size:** labelled as up to 50 participants ($ds \leqslant 50$), up to 100 participants ($ds \leqslant 100$) or over 100 participants ($ds > 100$). The results show that 13 studies were carried on smaller datasets (i.e. $ds \leqslant 50$), 24 studies carried on medium-sized datasets (i.e. $ds \leqslant 100$) and 14 studies carried on modestly larger datasets (i.e. $ds > 100$). However, seven of the studies carried on a medium-sized dataset and one study carried on a larger dataset attempted 3-way or even 4-way classification. Therefore, the group sizes of these studies are further reduced by the fact that the original dataset size needs to be divided into three or four groups, instead of the two groups used for binary classification.

We decided to use these numerical labels to classify the datasets, instead of assigning categories such as small or large, because even the largest dataset of the reviewed studies is relatively small when put into a machine learning context. All in all, there is a clear need for larger available datasets that are also balanced in terms of class and main risk factors. On larger datasets, it should be more straightforward to increase methodological rigour (e.g. by using CV, hold-out sets) and to seek for more active and systematic ways to prevent overfitting.

Strengths/Limitations (table 8: Clinical applicability)

In our view, a few desirable qualities should be present in AI research for AD, in order for it to be finally translatable into clinical practice. These are:

– **Spontaneous speech:** we consider spontaneous speech data to be more representative of real life spoken language. Although speech data obtained through non-spontaneous, constrained cognitive tasks present methodological advantages, we argue that spontaneous speech is desirable for cognitive monitoring due to its ubiquity, naturalness and relative ease of collection. Under this criterion, we seek not only to explore the advantages of using speech for cognitive screening, but also the suitability for continuous and longitudinal collection. 65% of the papers meet this criterion with this by using open question data (e.g. free episodic recalls, discourses prompted by a picture, conversational dialogues). The remaining papers rely on contrained data, obtained for example by recording the words produced in a fluency test.
– **Conversational speech:** similarly, we deem conversational speech to be more representative of real life spoken language than monologue speech. Here again we find a trade-off between naturalness and standardisation. While monologueas are easier to handle (by requiring fewer preprocessing steps) and may avoid potential confunding factors present in dialogues (e.g. relationships between speakers, conversational style, cultural norms surrounding doctor-patient conversations), some methods may take advantage of these very factors for cognitive screening as they enrich the cognitive mechanisms involved in the interaction [121]. Of the reviewed papers, only 18% report the use of dialogue (i.e. structured, semi-structured or conversational).
– **Automation:** most of the reviewed papers claim some degree of automation in their procedure, but looking closely, only 37% describe a fully (or nearly fully) automatic method, from transcription to classification. Another 37% describe a partially automatic procedure, frequently automating feature generation and/or classification steps, but with a manual transcription and/or manual feature set reduction. The rest describe methods that require manual interference at almost every stage, and were therefore deemed to not be automatic.
– **Content-independence:** this is desirable in order for successful methods to be equally successful when speech is elicited in different ways (i.e. with different tasks, which imply different content). 55% of the papers report procedures that do rely on content-related characteristics of speech, such as word content. The rest either rely solely on acoustic features or phoneme based transcribed features, unrelated to word content.
– **Transcription-free:** as mentioned above, ASR methods are an automatic alternative to manual transcription, but they are not free of constrains. Therefore, we consider transcription free approaches to offer a more relevant contribution to the clinical application of AI for AD detection. Under this criterion, 35% of the reviewed papers use a transcription-free approach, whereas the rest rely on either manual or ASR transcriptions.

Only two studies meet all five criteria with a "yes" [10, 84]. In our view, the field needs to further explore the use spontaneous speech (ideally conversational), and indeed we have observed renewed interest in its

use during the time span of this review, as AI becomes increasingly involved, as shown in Figure 1. Automation also needs to be pursued by trying to bridge the gaps where automation becomes challenging, namely, during transcription, as well as during feature generation and feature set reduction (i.e. feature selection and feature extraction). Seeking automation entails a complex trade-off, since there is clearly valuable information about a person's cognitive status reflected in the content of what they say, as well as how they say it and how they choose to structure it. In addition, not all linguistic features are content-dependent and metrics such as word frequency, vocabulary richness, repetitiveness and syntactic complexity are not linked to semantic content or meaning. However, processing language to obtain these metrics makes automation and generalisation more difficult, specially as regards non-English data. While content-related information can offer insights into the nature of the disease and its development, reviewing the potential for AI systems in terms of practical usefulness in clinical settings for cognitive health monitoring requires considerations of content-independent and transcription-free approaches due to their ease of implementation, successful performance and more straightforward generalisability.

Overall Conclusions

We have conducted the first systematic review on the potential application of interactive artificial intelligence methods to AD detection and progression monitoring using natural language processing and speech technology to extract "digital biomarkers" for ML modelling.

Given the somewhat surprising quantity and variety of studies we found, it seems reasonable to conclude that this is a very promising field, with potential to gradually introduce changes into clinical practice. Almost all studies report relatively high performance, despite the difficulties inherent to the type of data used and the heterogeneity of the methods. When compared to neuropsychological assessment methods, speech and language technology were found to be at least equally discriminative between different groups. It is worth noting that the most commonly used neuropsychological test, MMSE, has been criticised [72] due to its inherent biases and lack of sensitivity to subtle symptoms. In this context, interactive AI could offer the same or better performance as a screening method, with the additional advantages of being implemented automatically and, possibly, remotely. Notwithstanding, while most of the papers hereby reviewed highlight the potential of AI and ML methods, no actual translation into clinical practice has been achieved. One might speculate that this slow uptake, despite nearly 20 years of research in this field, is due to difficulties in attaining meaningful interdisciplinary cooperation among between AI/ML research experts and clinicians. We expect that the growing interest in and indeed adoption of AI/ML methods in medicine will provide the stimulus needed for effective translation to take place in this field as it has in others. Despite an unexpectedly high number of records found eligible to review (51), the field remains highly heterogeneous with respect to the available data and methodology. It is difficult to compare results on an equal footing when their conclusions are drawn from monologue, dialogue, spontaneous and non-spontaneous speech data. Similarly, different choices of processing units (varying from phoneme and syllable to a word, sentence or a longer instance) pose additional comparability challenges. Furthermore, while machine learning methodology is somewhat standardised through a wealth of open-source tools, the feature generation and feature set reduction procedures are not. Feature generation varies greatly, with the same feature falling into slightly different categories depending on the study. Consequently, abiding by a standard taxonomy like the one proposed by Voleti et al. [19], which we adapted in Table 1, becomes essential in order to make cross-study comparisons. Surprisingly, many studies do not report on their approach to feature set reduction, or do it very vaguely, giving less than enough detail for replication. To our knowledge, only one study [11] relies on standardised feature sets available to the research community, while all other articles extract and calculate speech and language indices in an *ad hoc*, non-consensual way.

Furthermore, although cross-validation is implemented in most publications as an evaluation technique, many studies proceed with feature set reduction outside a cross-validation setting. That is, both training and testing data are used to find the relevant features that will serve as input to the classifier input. Additionally, although it is standard practice to tune machine learning models using a preferred performance metric (i.e. *acc, EER, AUC, F1*), we must recognise the potential effect this might have on the reliability and generalisability of such models. If CV is done correctly (i.e. not optimizing hyper-parameter tuning within the test set of each fold), the models created in any given fold of the CV procedure are tested on unseen data, al-

though many studies do not provide this information. Barely any of the reviewed studies reported a hold-out set procedures or experiments on an entirely separate dataset, which would be the ideal scenario for robust model validation.

One of the reasons behind this lack of rigour is the size and variable quality of the datasets, which prevents adequate subsets to be generated while the size and integrity of the experimental groups is maintained. Consequently, we are confident that establishing certain standards on data and methodology will also increase the strictness of study evaluation. With regards to data type and availability, firstly, we should mention that data collection in this field is particularly difficult due to ethic constraints, due to the personally-identifying nature of speech data. Secondly, a benchmark dataset is essential to set the long overdue baselines of the field. Such baselines should not only refer to detection performance for SCI, MCI and AD classes, but also to regression models able to predict cognitive scores, which is repeatedly proposed but hardly ever done, and prediction of progression and risk. Thirdly, we note that conversational dialogue (i.e. natural dialogue) is an under-explored resource in this field. As noted before, although monologue data presents methodological advantages, dialogue data has the potential to offer richer results precisely due to factors that under certain methodological frameworks might be dismissed as confounds. That is, an AI system trained to evaluate speakers interaction, cultural norms and conversational styles has potential to be more versatile in monitoring cognitive health for different people, in different settings and at different times of the day. Furthermore, dialogue data could be easier and more natural to collect in real life (i.e. we spend part of our day interacting with somebody else), as well as more representative of a broader range psycholinguistic aspects such as alignment and entrainment at different linguistic levels [121], which might be relevant to AD detection.

With regards to methodology, we recommend a wider use of standardised feature sets, such as eGeMAPS [122], purposefully developed to detect physiological changes in voice production. Needless to say, other feature sets should also be built and tested systematically, for the field to move toward finding a golden standard. Further benefits of a consensual set of features entail the possibility of tracking those features "back to the brain", in order to find their neural substrate and hence contributing to knowledge of the neuropathology of AD.

In terms of aims and objectives, research suggests that embedded devices installed in the home to monitor patient health and safety may delay institutionalisation [123], and therefore more emphasis should be placed on the feasibility of remote evaluations. To this end, we propose that future research focuses on natural conversations, which are straightforward to collect passively, continuously and longitudinally in order to monitor cognitive health through an AI system. We also argue that focusing on cohorts already diagnosed with AD is no longer a relevant task for AI. As noted earlier, the majority of studies reviewed in this paper focus on diagnosis. We argue that emphasis should shift towards earlier stages of the disease, when pre-clinical unobserved changes start. Future research should therefore attempt to include healthy populations at higher risk of developing AD in larger scale longitudinal studies, as well as compare those populations to lower risk populations. There is good potential for interactive AI technology to contribute at those stages, given its increasingly ubiquitous presence in our lives, through wearable devices, smartphones, "smart homes", and "smart cities".

In addition, novel AI/ML digital biomarkers [124] could be used in combination with established biomarkers to target populations at risk of later dementia onset, as has already been proposed [101]. It needs to be emphasised that recorded data are considered personal data (i.e. with potential to identify a subject), with the ethical and regulatory hurdles this entails as regards data collection and analysis. We suggest that the field would benefit from revised ethics agreements to facilitate speech data collection, as well as from data sharing across institutions until datasets reach sufficient size to support complex machine learning structures and results are robust enough to encourage clinical applications. Increased collaboration between clinicians and AI experts should favour these developments.

Acknowledgements

We thank Marshall Dozier, the Academic Support Librarian for her help with the search queries and the PROSPERO protocol. This research was supported by the Medical Research Council (MRC), grant number MR/N013166/1.

1. Conflict of Interest/Disclosure Statement

The authors have no conflict of interest to report.

References

[1] American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[2] World Health Organization (2013) Mental health action plan 2013-2020. *WHO Library Cataloguing-in-Publication DataLibrary Cataloguing-in-Publication Data*, pages 1–44.

[3] Ross, GW, Cummings, JL, and Benson, DF (1990) Speech and language alterations in dementia syndromes: Characteristics and treatment. *Aphasiology*, 4(4):339–352.

[4] Watson, CM (1999) An analysis of trouble and repair in the natural conversations of people with dementia of the alzheimer's type. *Aphasiology*, 13(3):195–218.

[5] Bucks, RS, Singh, S, Cuerden, JM, and Wilcock, GK (2000) Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

[6] Luz, S (2017) Longitudinal Monitoring and Detection of Alzheimer's Type Dementia from Spontaneous Speech Data. In *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*, pages 45–46. IEEE.

[7] Fraser, KC, Lundholm Fors, K, Eckerstrom, M, Ohman, F, and Kokkinakis, D (2019) Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Front Aging Neurosci*, 11.

[8] Mirheidari, B, Blackburn, D, Walker, T, Reuber, M, and Christensen, H (2019) Dementia detection using automatic analysis of conversations. *Computer Speech and Language*, 53:65–79.

[9] Fraser, KC, Meltzer, JA, and Rudzicz, F (2016) Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

[10] Luz, S, la Fuente, SD, and Albert, P (2018) A method for analysis of patient speech in dialogue for dementia detection. In Kokkinakis, D, editor, *Procs. of LREC'18*, Paris, France. ELRA.

[11] Haider, F, De La Fuente Garcia, S, and Luz, S (2019) An assessment of paralinguistic acoustic features for detection of Alzheimer's Dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*.

[12] Mirheidari, B, Blackburn, D, O'Malley, R, Walker, T, Venneri, A, Reuber, M, and Christensen, H (2019) Computational Cognitive Assessment: Investigating the Use of an Intelligent Virtual Agent for the Detection of Early Signs of Dementia. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2732–2736.

[13] Analytics, C (2016) Endnote x8: Philadelphia. *United States*.

[14] Richardson, WS, Wilson, MC, Nishikawa, J, Hayward, RS, et al. (1995) The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–3.

[15] Leeflang, MM, Deeks, JJ, Gatsonis, C, and Bossuyt, PM (2008) Systematic reviews of diagnostic test accuracy. *Annals of internal medicine*, 149(12):889–897.

[16] Whiting, PF, Rutjes, AW, Westwood, ME, Mallett, S, Deeks, JJ, Reitsma, JB, Leeflang, MM, Sterne, JA, and Bossuyt, PM (2011) Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 155 (8):529–536.

[17] Wolff, RF, Moons, KG, Riley, RD, Whiting, PF, Westwood, M, Collins, GS, Reitsma, JB, Kleijnen, J, and Mallett, S (2019) Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58.

[18] Higgins, JP and Altman, DG (2008) Assessing risk of bias in included studies. *Cochrane handbook for systematic reviews of interventions: Cochrane book series*, pages 187–241.

[19] Voleti, RNU, Liss, J, and Berisha, V (2019) A review of automated speech and language features for assessment of cognition and thought disorders. *IEEE Journal of Selected Topics in Signal Processing*.

[20] Tausczik, YR and Pennebaker, JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.

[21] Yngve, VH (1960) A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

[22] Frazier, L (1985) *Syntactic Complexity*. Cambridge University Press, Cambridge, UK.

[23] Mikolov, T, Chen, K, Corrado, G, and Dean, J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[24] Blei, DM, Ng, AY, and Jordan, MI (2003) Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[25] Davis, S and Mermelstein, P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

[26] Folstein, MF, Folstein, SE, and McHugh, PR (1975) "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

[27] Nasreddine, ZS, Phillips, NA, Bédirian, V, Charbonneau, S, Whitehead, V, Collin, I, Cummings, JL, and Chertkow, H (2005) The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

[28] Cole, M, Dastoor, D, and Simpson, T (2015) *Hierarchic Dementia Scale - Revised: Instruction Manual*. Dementia Training Study Centre.

[29] Morris, JC (1991) The clinical dementia rating (cdr): Current version and. *Young*, 41:1588–1592.

[30] Freedman, M, Leach, L, Kaplan, E, Shulman, K, and Delis, DC (1994) *Clock drawing: A neuropsychological analysis*. Oxford University Press, USA.

[31] Rosen, WG, Mohs, RC, and Davis, KL (1984) A new rating scale for alzheimer's disease. *The American journal of psychiatry*.

[32] Joanette, Y, Ska, B, Poissant, A, Belleville, S, Bellavance, A, Gauthier, S, Gauvreau, D, Lecours, A, and Peretz, I (1995) Évaluation neuropsychologique et profils cognitifs des démences de type alzheimer: dissociations transversales et longitudinales. *Dans F. Eustache et A. Agniel (Éds.), Neuropsychologie clinique des démences*, pages 91–106.

[33] Brodaty, H, Pond, D, Kemp, NM, Luscombe, G, Harding, L, Berman, K, and Huppert, FA (2002) The gpcog: a new screening test for dementia designed for general practice. *Journal of the American Geriatrics Society*, 50(3):530–534.

[34] Peña, MM, Carrasco, PM, Luque, ML, and García, AIR (2012) Evaluación y diagnóstico del deterioro cognitivo leve. *Revista de logopedia, foniatría y audiología*, 32(2):47–56.

[35] Reisberg, B, Ferris, SH, De Leon, M, Crook, T, et al. (1988) Global deterioration scale (gds). *Psychopharmacol Bull*, 24 (4):661–3.

[36] Wallace, M, Shelkey, M, et al. (2007) Katz index of independence in activities of daily living (adl). *Urol Nurs*, 27(1):93–94.

[37] Graf, C (2009) The lawton instrumental activities of daily living (iadl) scale. *The gerontologist*, 9(3):179–186.

[38] Petersen, RC, Smith, GE, Waring, SC, Ivnik, RJ, Kokmen, E, and Tangelos, EG (1997) Aging, memory, and mild cognitive impairment. *International psychogeriatrics*, 9(S1):65–69.

[39] McKhann, G, Drachman, D, Folstein, M, Katzman, R, Price, D, and Stadlan, EM (1984) Clinical diagnosis of alzheimer's disease: Report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939.

[40] Schmidtke, K, Pohlmann, S, and Metternich, B (2008) The syndrome of functional memory disorder: definition, etiology, and natural course. *The American Journal of Geriatric Psychiatry*, 16(12):981–988.

[41] Benton, A, Hamsher, K, and Sivan, A (1976) Multilingual aphasia examination. iowa city. *University of Iowa*.

[42] Goodglass, H and Kaplan, E (1983) *The Boston Diagnostic Aphasia Examination*. Lea & Febinger, Philadelphia.

[43] Sadeghian, R, Schaffer, JD, and Zahorian, SA (2017) Speech Processing Approach for Diagnosing Dementia in an Early Stage. pages 0–4.

[44] Wechsler, D (1997) *WAIS-III: Wechsler adult intelligence scale (third edition manual)*. Psychological Corporation.

[45] Wechsler, D (1997) *WMS-III: Wechsler memory scale administration and scoring manual (third edition)*. Psychological Corporation.

[46] Nelson, HE and Willison, J (1991) *National adult reading test (NART)*. Nfer-Nelson Windsor.

[47] Bayles, KA, Tomoeda, CK, and Dharmaperwira-Prins, R (1993) *ABCD: Arizona battery for communication disorders of dementia*. Canyonlands Publishing.

[48] Darley, FL, Aronson, AE, and Brown, JR (1975) *Motor speech disorders*. Saunders.

[49] Mirzaei, S, El Yacoubi, M, Garcia-Salicetti, S, Boudy, J, Kahindo, C, Cristancho-Lacroix, V, Kerherve, H, and Rigaud, AS (2018) Two-Stage Feature Selection of Voice Parameters for Early Alzheimer's Disease Prediction. *Irbm*, 39(6):430–435.

[50] Becker, JT, Boiler, F, Lopez, OL, Saxton, J, and McGonigle, KL (1994) The Natural History of Alzheimer's Disease. *Archives of Neurology*, 51(6):585.

[51] MacWhinney, B (2019) Understanding spoken language through talkbank. *Behavior research methods*, 51(4):1919–1927.

[52] Gósy, M (2013) Bea–a multifunctional hungarian spoken language database. *Phonetician*, 105:50–61.

[53] Wallin, A, Nordlund, A, Jonsson, M, Lind, K, Edman, Å, Göthlin, M, Stålhammar, J, Eckerström, M, Kern, S, Börjesson-Hanson, A, et al. (2016) The gothenburg mci study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism*, 36(1):114–131.

[54] Pope, C and Davis, BH (2011) Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

[55] Mirheidari, B, Blackburn, DJ, Harkness, K, Walker, T, Venneri, A, Reuber, M, and Christensen, H (2017) An avatar-based system for identifying individuals likely to develop dementia. In *Interspeech 2017*, pages 3147–3151. ISCA.

[56] Moher, D, Liberati, A, Tetzlaff, J, and Altman, DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

[57] Bertola, L, Mota, NB, Copelli, M, Rivero, T, Diniz, BS, Romano-Silva, MA, Ribeiro, S, and Malloy-Diniz, LF (2014) Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front Aging Neurosci*, 6:185.

[58] Konig, A, Satt, A, Sorin, A, Hoory, R, Derreumaux, A, David, R, and Robert, P (2018) Use of Speech Analyses within a Mobile Application for the Assessment of Cognitive Impairment in Elderly People. *Curr Alzheimer Res*, 15(2):120–129.

[59] Lundholm Fors, K, Fraser, KC, and Kokkinakis, D (2018) Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. *Stud Health Technol Inform*, 247:705–709.

[60] Martinez de Lizarduy, U, Calvo Salomon, P, Gomez Vilda, P, Ecay Torres, M, and Lopez de Ipina, K (2017) ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, 4(1).

[61] Satt, A, Sorin, A, Toledo-Ronen, O, Barkan, O, Kompatsiaris, I, Kokonozi, A, and Tsolaki, M (2013) Evaluation of speech-based protocol for detection of early-stage dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August):1692–1696.

[62] Egas López, JV, Tóth, L, Hoffmann, I, Kálmán, J, Pákáski, M, and Gosztolya, G (2019) Assessing Alzheimer's Disease from Speech Using the i-vector Approach. In *International Conference on Speech and Computer*, pages 289–298. Springer.

[63] Gosztolya, G, Vincze, V, Tóth, L, Pakaski, M, Kalman, J, and Hoffmann, I (2019) Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech and Language*, 53:181–197.

[64] Kato, S, Endo, H, Homma, A, Sakuma, T, and Watanabe, K (2013) Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2013:5813–6.

[65] Nasrolahzadeh, M, Mohammadpoory, Z, and Haddadnia, J (2018) Higher-order spectral analysis of spontaneous speech signals in Alzheimer's disease. *Cognitive Neurodynamics*, 12 (6):583–596.

[66] Thomas, C, Keselj, V, Cercone, N, Rockwood, K, and Asp, E (2005) Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, 3(February):1569–1574.

[67] Clark, DG, McLaughlin, PM, Woo, E, Hwang, K, Hurtz, S, Ramirez, L, Eastman, J, Dukes, RM, Kapur, P, DeRamus, TP, and Apostolova, LG (2016) Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 2:113–122.

[68] Weiner, J and Schultz, T (2016) Detection of Intra-Personal Development of Cognitive Impairment From Conversational Speech. In *Speech Communication; 12. ITG Symposium*, pages 1–5.

[69] Duong, A, Giroux, F, Tardif, A, and Ska, B (2005) The heterogeneity of picture-supported narratives in {Alzheimer}'s disease. *Brain and language*, 93(2):173–184.

[70] Prud'hommeaux, ET and Roark, B (2015) Graph-based word alignment for clinical language evaluation. *Comput. Linguist.*, 41(4):549–578.

[71] Shinkawa, K, Kosugi, A, Nishimura, M, Nemoto, M, Nemoto, K, Takeuchi, T, Numata, Y, Watanabe, R, Tsukada, E, Ota, M, Higashi, S, Arai, T, and Yamada, Y (2019) Multimodal Behavior Analysis Towards Detecting Mild Cognitive Impairment: Preliminary Results on Gait and Speech. *Stud Health Technol Inform*, 264:343–347.

[72] Carnero-Pardo, C (2014) Should the mini-mental state examination be retired? *Neurología (English Edition)*, 29(8):473–481.

[73] Prud'Hommeaux, ET and Roark, B (2011) Alignment of spoken narratives for automated neuropsychological assessment. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings*, pages 484–489.

[74] Weiner, J, Herff, C, and Schultz, T (2016) Speech-Based Detection of Alzheimer's Disease in Conversational German. In *17th Annual Conference of the International Speech Communication Association*, pages 1938–1942. ISBN 2308-457X 978-1-5108-3313-5.

[75] Orimaye, SO, Wong, JSM, Golden, KJ, Wong, CP, and Soyiri, IN (2017) Predicting probable {Alzheimer}'s disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18 (1):34.

[76] Rochford, I, Rapcan, V, D'Arcy, S, and Reilly, RB (2012) Dynamic minimum pause threshold estimation for speech analysis in studies of cognitive function in ageing. *Conf Proc IEEE Eng Med Biol Soc*, 2012:3700–3703.

[77] Tröger, J, Linz, N, Alexandersson, J, König, A, and Robert, P (2017) Automated speech-based screening for alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, Barcelona, Spain. ACM.

[78] Roark, B, Mithcell, M, Hosom, J, Hollingshead, K, and Kaye, J (2011) Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *The New England journal of medicine*, 19(7):2081–2090.

[79] Lopez-de Ipiña, K, Alonso-Hernandez, JB, Sole-Casals, J, Travieso-Gonzalez, CM, Ezeiza, A, Faundez-Zanuy, M, Calvo, PM, and Beitia, B (2015) Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer's disease. *Neurocomputing*, 150(PB):392–401.

[80] Lopez-de Ipiña, K, Alonso, JB, Solé-Casals, J, Barroso, N, Henriquez, P, Faundez-Zanuy, M, Travieso, CM, Ecay-Torres, M, Martinez-Lage, P, and Egiraun, H (2015) On Automatic Diagnosis of Alzheimer's Disease based on Spontaneous Speech Analysis and Emotional Temperature. *Cognitive Computation*, 7(1):44–55.

[81] Mielke, MM, Vemuri, P, and Rocca, WA (2014) Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. *Clinical epidemiology*, 6:37.

[82] Chien, Y, Hong, S, Cheah, W, Fu, L, and Chang, Y (2018) An Assessment System for Alzheimer's Disease Based on Speech Using a Novel Feature Sequence Design and Recurrent Neural Network. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3289–3294.

[83] Espinoza-Cuadros, F, Garcia-Zamora, MA, Torres-Boza, D, Ferrer-Riesgo, CA, Montero-Benavides, A, Gonzalez-Moreira, E, and Hernandez-Gómez, LA (2014) A spoken language database for research on moderate cognitive impairment: Design and preliminary analysis. In Navarro Mesa, JL, Ortega, A, Teixeira, A, Hernández Pérez, E, Quintana Morales, P, Ravelo García, A, Guerra Moreno, I, and Toledano, DT, editors, *Advances in Speech and Language Technologies for Iberian Languages, IberSpeech 2014*, volume 8854, pages 219–228, Cham. Springer International Publishing.

[84] Khodabakhsh, A and Demiroğlu, C (2015) Analysis of speech-based measures for detecting and monitoring Alzheimer's disease. *Methods Mol Biol*, 1246:159–173.

[85] Mirheidari, B, Blackburn, D, Walker, T, Venneri, A, Reuber, M, and Christensen, H (2018) Detecting Signs of Dementia Using Word Vector Representations. In *Interspeech*, pages 1893–1897.

[86] Mirheidari, B, Blackburn, D, Harkness, K, Walker, T, Venneri, A, Reuber, M, and Christensen, H (2017) Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of {Alzheimer}'s Disease*, 58 (2):373–387.

[87] Fraser, KC, Lundholm Fors, K, and Kokkinakis, D (2019) Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139.

[88] Rentoumi, V, Paliouras, G, Danasi, E, Arfani, D, Fragkopoulou, K, Varlokosta, S, and Papadatos, S (2017) Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 33–38.

[89] Gonzalez-Moreira, E, Torres-Boza, D, Kairuz, H, Ferrer, C, Garcia-Zamora, M, Espinoza-Cuadros, F, and Hernandez-Gómez, L (2015) Automatic prosodic analysis to identify mild dementia. *BioMed Research International*.

[90] Dos Santos, LB, Corrêa, EA, Oliveira, ON, Amancio, DR, Mansur, LL, and Aluísio, SM (2017) Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pages 1284–1296. Association for Computational Linguistics (ACL).

[91] Beltrami, D, Calzà, L, Gagliardi, G, Ghidoni, E, Marcello, N, Favretti, RR, and Tamburini, F (2016) Automatic Identification of Mild Cognitive Impairment through the Analysis of Italian Spontaneous Speech Productions. In *Proceedings of*

the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 2086–2093.

[92] Guo, Z, Ling, Z, and Li, Y (2019) Detecting Alzheimer's Disease from Continuous Speech Using Language Models. *Journal of Alzheimers Disease*, 70(4):1163–1174.

[93] D'Arcy, S, Rapcan, V, Penard, N, Morris, ME, Robertson, IH, and Reilly, RB (2008) Speech as a means of monitoring cognitive function of elderly speakers. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, number January 2008, pages 2230–2233.

[94] Tanaka, H, Adachi, H, Ukita, N, Ikeda, M, Kazui, H, Kudo, T, and Nakamura, S (2017) Detecting Dementia Through Interactive Computer Avatars. *IEEE J Transl Eng Health Med*, 5:2200111.

[95] Ben Ammar, R and Ben Ayed, Y (2018) Speech Processing for Early Alzheimer Disease Diagnosis: Machine Learning Based Approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.

[96] Tröger, J, Linz, N, König, A, Robert, P, and Alexandersson, J. Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment, (2018).

[97] Yu, B, Williamson, JB, Mundt, JC, and Quatieri, TF (2018) Speech-Based Automated Cognitive Impairment Detection From Remotely-Collected Cognitive Test Audio. *IEEE Access*, 6:40494–40505.

[98] Guinn, C, Singer, B, and Habash, A (2014) A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pages 98–103.

[99] Tóth, L, Hoffmann, I, Gosztolyac, G, Vinczec, V, Szatlóczkid, S, Bánrétib, Z, Pákáskid, M, and Kálmán, J (2018) A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. 2050: 130–138.

[100] Meilan, JJ, Martinez-Sanchez, F, Carro, J, Lopez, DE, Millian-Morell, L, and Arana, JM (2014) Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord*, 37(5-6):327–334.

[101] de la Fuente, S, Ritchie, C, and Luz, S (2019) Protocol for a conversation-based analysis study: Prevent-ED investigates dialogue features that may help predict dementia onset in later life. *BMJ Open*, 9(3).

[102] Konig, A, Satt, A, Sorin, A, Hoory, R, Toledo-Ronen, O, Derreumaux, A, Manera, V, Verhey, F, Aalten, P, Robert, PH, and David, R (2015) Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)*, 1(1):112–124.

[103] Danso, S, Terrera, G, Luz, S, and Ritchie, C (2019) Application of big data and artificial intelligence technologies to dementia prevention research: an opportunity for low-and-middle-income countries. *Journal of Global Health*.

[104] Madikeri, S, Dey, S, Motlicek, P, and Ferras, M. Implementation of the standard i-vector system for the kaldi speech recognition toolkit. Technical report, Idiap, (2016).

[105] Biber, D, Connor, U, and Upton, T (2007) Discourse on the move. *Using corpus analysis to describe discourse structure*.

[106] Riley, KP, Snowdon, DA, Desrosiers, MF, and Markesbery, WR (2005) Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study. *Neurobiology of aging*, 26(3):341–347.

[107] Resnik, P (1992) Left-corner parsing and psychological plausibility. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pages 191–197. Association for Computational Linguistics.

[108] Pakhomov, S, Chacon, D, Wicklund, M, and Gundel, J (2011) Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behavior Research Methods*, 43(1):136–144.

[109] 't Hart, J (1981) Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3):811–821.

[110] Eyben, F, Wöllmer, M, and Schuller, B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

[111] Linz, N, Tröger, J, Alexandersson, J, Wolters, M, König, A, and Robert, P (2017) Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728. IEEE.

[112] Al-Hameed, S, Benaissa, M, and Christensen, H (2017) Detecting and predicting alzheimer's disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pages 57–61.

[113] Pou-Prom, C and Rudzicz, F (2018) Learning multiview embeddings for assessing dementia. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2812–2817.

[114] Yancheva, M, Fraser, KC, and Rudzicz, F (2015) Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 134–139.

[115] Luz, S, Haider, F, de la Fuente, S, Fromm, D, and MacWhinney, B (2020) Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*.

[116] Haider, F, de la Fuente Garcia, S, Albert, P, and Luz, S (2020) Affective speech for alzheimer's dementia recognition. In *LREC 2020 Workshop: Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID-3 @ LREC). Marseille, France*, pages 191–197. European Language Resources Association.

[117] Horley, K, Reid, A, and Burnham, D (2010) Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*.

[118] Pakhomov, SVS and Hemmy, LS (2014) A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*, 55(1):97–106.

[119] Wang, F, Kaushal, R, and Khullar, D (2019) Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Annals of Internal Medicine*, 172(1): 59–60.

[120] Molnar, C (2019) *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

[121] Pickering, MJ and Garrod, S (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27 (02).

[122] Eyben, F, Scherer, KR, Schuller, BW, Sundberg, J, André, E, Busso, C, Devillers, LY, Epps, J, Laukka, P, Narayanan, SS, et al. (2016) The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

[123] Fredericks, EM, Bowers, KM, Price, KA, and Hariri, RH (2018) Cal: A smart home environment for monitoring cognitive decline. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 1500–1506. IEEE.

[124] Coravos, A, Khozin, S, and Mandl, KD (2019) Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digital Medicine*, 2(11):1–5.

[125] Ekberg, K and Reuber, M (2015) Can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions? *Communication & medicine*, 12(1):13.

[126] Rockwood, K (2004) The acadie study: Does donepezil meet the expectations of treatment? *Canadian Alzheimer Disease Review*, page 13.

[127] Weiner, J, Frankenberg, C, Telaar, D, Wendelstein, B, Schröder, J, and Schultz, T (2016) Towards automatic transcription of ILSE – an interdisciplinary longitudinal study of adult development and aging. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 718–725.

[128] Brian, M (2000) The childes project: Tools for analyzing talk. *Mahwah, NJ & London: Lawrence Erlbaum*.

[129] Lewis, Z (1995) *Walt Disney's Cinderella*. Random House Disney.

[130] Ten, HP. Doing conversation analysis, (2007).

[131] Mak, MW and Yu, HB (2014) A study of voice activity detection techniques for nist speaker recognition evaluations. *Computer Speech & Language*, 28(1):295–313.

[132] Reynolds, D (2009) *Universal Background Models*, pages 1349–1352. Springer US, Boston, MA.

[133] Pears, R, Finlay, J, and Connor, AM (2014) Synthetic minority over-sampling technique (smote) for predicting software build outcomes. *arXiv preprint arXiv:1407.2330*.

Supplementary material

Keys to table interpretability

The following conventions and acronyms were adopted when reviewing the research articles.

When a conference paper was later extended and published in a different venue, only the latter publication was included. Percentages were rounded to the nearest decimal place. Education is expressed in average years, same as age, unless otherwise specified. For the purposes of this review we do not distinguish between acoustic and paralinguistic features, and the tables we will always designate these features as "acoustic". Where the type of classifier or model used is not specified on the table, this means it was not reported in the paper reviewed. Since "data balance" can be understood in, at least, three different ways, we created the following acronyms to standardise and simplify the description of this characteristic on the tables:

- **dataset Feature Balance (FB):** where the "F" is replaced for the initial of each particular feature. This refers to whether a certain feature, e.g. gender, is evenly distributed in the dataset as a whole. Consider, for instance, a dataset with 100 participants, 60 healthy controls (HC), 30 of whom are female, and 40 AD participants, 30 of whom are female. As regards gender, this is indicated in the table as "No-GB" (no gender balance) because the ratio of female to male participants in this dataset is 60:40. Regarding class, this dataset would also be "No-CB" (no class balance), because the ratio of HC to AD is also 60:40A. Age and education are reported as a between class features only.
- **Within-Classes Feature Balance (WCFB):** This indicates whether a certain feature, such as gender, is evenly distributed *within* each class. In the above described sample dataset, this would be indicate as "HC: WCGB, AD: no-WCGB", because the gender ratio is 30:30 in HC, but 30:10 in AD. This type of balance makes sense for gender, since it is a category within the classes. However, age and education are generally reported as group averages, and hence it is not possible to report their balance within class.
- **Between-Classes Feature Balance (BCB: F):** This denotes whether a feature is evenly distributed *across* experimental classes. Our example dataset would be described as "BCB: no-G", because the number of female is indeed balanced (30 in both groups), but the number of males is not (30 in HC vs 10 in AD). Again, his type of balance makes sense for gender, since it is a category within the classes. However, age and education are generally reported as group averages, and hence balance between classes is equivalent to their balance across the dataset. Since they are reported to control for their potential confounding effect between groups, we will report their between-class balance.

Gender balance is reported as GB, WCGB, BCB: G, or, alternatively, no-GB, no-WCGB, BCB: no-G. Class balance is reported as CB or no-CB. Age and education, with respect to group average are reported within BCB: A, E or BCB: no-A, no-E. No-BCB indicates that none of the features are balanced between classes, whereas BCB (without further specification) indicates that all three are. Lastly, some studies report the number of speech or text samples as well as the number of participants per group and then take only one of those figures for analysis. In these cases class balance will be indicated followed by an *m* or an *n* depending on whether the comparison groups are balanced in terms of samples or participants, respectively. For example, CB*m* would indicate class balance based on number of samples per class, whereas no-CB*n* would indicate class imbalance in terms of number participants per class, both of which could coexist in the same study. A similar logic applies to other types of balance (e.g. no-WCGB*m*, BCB*m*: no-A, G, no-E). When it is unspecified whether the analyses and reported results are based on number of participants or number of samples, this will be indicated as "unclear".

Abbreviations and Acronyms

Table 2
Diagnostic abbreviations

Diagnostic Groups	
AD	Alzheimer's Disease or Dementia
CI	Cognitive Impairment (unspecified)
HC	Healthy Controls
MCI	Mild Cognitive Impairment
SCI	Subjective Cognitive Impairment

Note: For the purpose of this review, labels such as "normal elderly" or "cognitively normal" are noted as HC; dementia groups as AD. For the pre-clinical stage, Subjective Memory Loss (SML) is equated to SCI. CI refers to the symptomatic group where no official diagnosis term is stated.

Table 3
General textual abbreviations

General terms and noun phrases	
ast	assessment
avail	available
B/L	baseline
corr	correlation
demogr	demographics
ds	dataset
ft	features
ibid	see previous footnote on same dataset/paper/topic
incl	included
info	information
lex	lexical
m	number of samples
meas	measurement (s)
n	number of participants
NI	Neuroimaging
pp	participants
rec	recording
repr	representation
S/N	average sentences per narrative or sample
seg	segmentation
syl	syllable
tr	transcript
utt	utterance
w/	with
w/o	without
W/S	average words per sentence or sample

Table 4
Methods and metrics

Methods	
ADR	Active Data Representation
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CV	Cross-validation
DR	Data representation
DT	Decision Trees
GC	Gaussian Classifier
GNB	Gaussian Naive Bayes
HOS	Higher Order Spectral (analysis)
IG	Information Gain
k-NN	*k*-Nearest Neighbour
LDA	Linear Discriminant Analysis
LASSO (LR)	Least Absolute Shrinkage and Selection Operator
LM	Language Model
LR	Logistic Regression
LOO (CV)	Leave-one-out
LPO (CV)	Leave-pair-out
LSA	Latent Semantic Analysis
LSTM (RNN)	Long-short-term-memory
MLP	Multi-layer Perceptron
NN	Neural Network
PCA	Principal Components Analysis
RBF (SVM)	Radial Basis Function (kernel)
RF	Random Forest
RFE	Recursive Feature Elimination
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
Metrics	
ACC	Accuracy
AUC	Area Under the Curve
CER	Classification Error Rate
EER	Equal Error Rate
FA	False Alarms
MLU	mean length of utterance
N	noun frequency
pc	Precision
rc	Recall
sp	Specificity
ss	Sensitivity
ROC	Receiving operating characteristic (curve)
TP	True Positives
UAR	Unweighted Average Recall
V	verb frequency

2. SPICMO (PICOS) table

The first table is based on the PICOS design, widely used in the clinical field. Its columns are:

- **Population**: total number of participants followed by number of participants per group (always starting with the less impaired group). Average demographic figures (i.e. age, education, MMSE) follow the same order.
- **Interventions**: assessments that participants underwent as part of the study. These are usually either cognitive or full clinical assessments, recorded speech tasks and written tasks.
- **Comparison groups**: different stages of cognitive impairment of the study participants conform the groups to be compared. These are Healthy Controls (HC), Subjective Cognitive Impairment (SCI), Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD) and Cognitive Impairment (CI), when unspecified. This terminology is not standardised across publications, but we have standardised it for the purpose of this review. Hence, for instance, normal controls (NC), healthy elderly (HE), Subjective Memory Complaints (SMC) or Dementia (e.g. Alzheimer's Type Dementia) are hereby equated to HC, SCI and AD, respectively.
- **Outcomes of interest**: detection, prediction or discrimination performance of the method used in an article. This mostly includes classification metrics, such as overall accuracy, sensitivity and specificity.
- **Study aim/design**: most frequently, automatic detection of a target group when compared to a healthy one, or automatic discrimination between different stages of target groups. It also includes the main design, i.e. text vs. speech, narrative vs. monologue.

We have extended this by adding a column on methods, an essential part of this review:

- **Methodology**: brief overview of the approach for feature generation (i.e. acoustic analysis or natural language processing), as well as the approach for feature set reduction (when reported). These are feature selection (i.e. filtering or wrapping) and feature extraction (i.e. combination or transformation of original features, e.g. PCA, LSA, ADR). This section also mentions the machine learning task used in the paper (i.e. machine learning task).

Lastly, we considered it to be more intuitive for this review to have information about **S**tudy aim at the beginning, and therefore, the conventional order of the columns has been shifted, yielding SPICMO (study aim, population, intervention, comparisons, methodology and outcomes) as a result.

Table 5: SPICMO (PICOS) table

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Beltrami et al. [91]	Automatic detection of MCI based on acoustic and linguistic fts from narrative speech data.	39 pps: 20 HC, 19 MCI. Aged range: 50-75 years old; educ: high school/university. Italy.	Cognitive ast. **Speech task:** narratives picture description, working day, last dream.	HC and MCI. Based on cognitive ast (MMSE, MoCA, GPCog, CDT, VF, CANTAB-PAL)	Acoustic and linguistic analysis of speech for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 76.9% accuracy (picture task) ($F_1 = 78.1\%$).
Ben Ammar and Ben Ayed [95]	Automatic detection of AD based on linguistic fts from narrative speech data.	484 samples: 242 HC, AD; unreported pps from Pitt[1]. Age > 44; educ > 7; MMSE > 10. USA.	Clinical ast. **Speech task:** narrative picture description (Cookie Theft).	HC and MCI. Based on cognitive ast (i.e. MMSE).	Audio enhancement, linguistic analysis for ft extraction, ML for ft selection and group classification.	Detection performance: 79% accuracy.
Bertola et al. [57]	Discrimination between HC, MCI and AD using graph analysis on word sequences obtained from SVF data.	100 pps: 25 HC, aMCI, a+mdMCI, AD. Median age: 76, 76 (MCI), 79; educ: 4, 4, 4; MMSE: 27, 25, 20. Brazil.	Clinical ast (i.e. medical, cognitive). **Speech task:** recorded SVF answers (animals).	HC, amnesic single/multiple domain (a/+mdMCI), and AD. Based on cognitive ast (Katz, Lawton, MMSE).	Graph analysis for ft extraction from word sequences, statistics for ft selection (incl SVF scores), ML for group classification.	Discrimination performance: $AUC = 0.68$ HC-MCI, $AUC = 0.73$ MCI-AD, $AUC = 0.88$ HC-AD (graph attributes only).
Chien et al. [82]	Automatic detection of AD based on acoustic fts from narrative speech and SVF data.	60 pps: 30 HC, 30 AD from Mandarin_Lu[2]. 150 speech samples, demographics unreported. China, Taiwan.	**Speech task:** recorded SVF answers (fruits, locations) and narrative picture description.	HCand AD. Unreported criteria.	Manual acoustic analysis for ft extraction, ML for group classification.	Detection performance: $AUC = 0.95$.
Clark et al. [67]	Automatic prediction of MCI conversion to AD from automatic SVF scores combined with neuroimaging data.	107 pps: 83 MCI-non (46F), 24 MCI-con (15F). Avg age: 68.7, 73.8; educ: 16, 16; MMSE: 27.9, 25.1. USA.	Cognitive ast, brain MRI. **Speech task:** recorded SVF (vegetables, animals) and OVF (letters F, A, S).	MCI-non and MCI-con (upon conversion to AD). Based on Petersen criteria (incl. MMSE, CDR).	Automatic scoring of verbal fluency answers (electronically transcribed) and neuroimaging scores for ft extraction, ML for group classification.	Conversion prediction performance (at 4-year follow-up): $AUC = 0.872$ with automatic scores.
D'Arcy et al. [93]	Detection of probable CI based on manually and automatically extracted temporal speech fts.	87 pps: 50 HC 37 probable CI. Age range: 62-92, 62%F, MMSE over 24 (excl severe CI). Ireland.	Cognitive ast . **Speech task**: narrative picture descriptions, Scribe[3], SVF (animals), word list, Heidi passage.	HC (MMSE > 27) and probable CI (MMSE $\leqslant$ 27).	Manual and automatic temporal analysis of speech for ft extraction (syntactic, acoustic), ML for group classification.	Detection performance: 76.74% accuracy (manual approach far superior than automatic).Vowel 17% longer in probable CI.
Dos Santos et al. [90]	Automatic detection of MCI based on speech fts from transcribed narratives in English and Portuguese.	86 Pitt[4] pps. USA. 40 Cinderella[5] pps. USA. 43 ABCD[6] pps. Brazil.	Cognitive ast. **Speech task:** picture description (Pitt), story retelling (Cinderella), recall (ABCD).	HC and MCI. Based on cognitive ast (Pitt), Petersen criteria (Cinderella), clinical diagnosis (ABCD).	Topological network and linguistic analysis for ft extraction, ML for group classification.	Detection performance: 65% Pitt accuracy, 65% Cinderella, 75% ABCD.
Duong et al. [69]	Description of discourse patterns and heterogeneity in AD based on transcribed narrative speech.	99 pps: 53 HC (40F), 46 AD (39F). Avg age: 73.8, 74.3; educ: 10.2, 8.3. Canada.	Cognitive ast. **Speech task:** P1 and P7 narrative picture descriptions from PENO[7].	HC (NE: normal elderly) and AD. Based on NINCDS-ADRDA criteria [39].	Discourse analysis of the transcribed narratives, cluster analysis to group pps with similar discourse patterns.	Clusters inconclusive for prototypical AD discourse (heterogeneity). 4 different discourse patterns for P1 and 5 for P7.
Egas López et al. [62]	Discrimination between HC, MCI and AD using the i-vector approach on acoustic fts from narrative speech.	75 pps: 25 HC, MCI, AD. Avg age: 73.96, 72.4, 70.72; educ: 10.76, 10.84, 12.08; MMSE: 29.24, 27.16, 23.92.	Cognitive ast. **Speech task:** narrative of previous day, immediate and delayed recall of 2 short films.	HC, MCI and AD. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog)	Acoustic analysis of speech recordings ft extraction, i-vector approach for dimensionality reduction, ML for group classification.	Discrimination performance: 56% accuracy and $F_1 = 78.4\%$, all tasks. $F_1 = 79.2\%$ immediate recall only.

[1] This paper reports the number of speech samples they used, but not to how many pps they belonged to.
[2] Mandarin_Lu corpus is hosted within DementiaBank
[3] 5 pictures from an in-house task (Picture Taboo). Scribe consists of sentences designed to cover English language phones.
[4] 86 pps: 43 HC (20F), MCI (16F). Avg age: 64.1, 69.3; educ over 7, MMSE over 10.
[5] 40pps: 20 HC (16F), MCI (14F). Avg age: 74.8, 73.3; educ: 11.4, 10.8.
[6] 43 pps: 20 HC, 23 MCI. Avg age: 61, 72; educ: 16, 13.3.
[7] PENO is a cognitive battery in French [32]. "Bank robbery" and "Car accident" are language subtests from this battery.

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Espinoza-Cuadros et al. [83]	Automatic detection of MCI based on speech fts from interviews and narratives.	19 pps: 11 HC (6F), 8 MCI (2F). Avg age: 78.9, 80.3; educ: 8, 5	**Speech task:** structured interview recorded from MEC[8] and a reading short passage.[9].	HC and MCI. Unreported criteria.	Acoustic analysis of speech recordings for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 78.9% accuracy with seven prosodic features from the passage reading task.
Fraser et al. [7]	Automatic detection of MCI based on multi- modal fts from language tasks (audio, text, eye-tracking).	55 pps: 29 HC (21F), 26 MCI (14F). Avg age: 67.8, 70.6; educ: 13.3, 14.3; MMSE: 29.6, 28.2.	Cognitive ast. **Speech task:** picture description (Cookie Theft), read short text[10] aloud and in silence.	HC and MCI. Based on Petersen criteria. Gothemburg MCI Study.	Multi-modal approach for ft extraction, cascaded ML for group classification (ft, mode, task and session).	Detection performance: 83% accuracy (AUC$=0.88$), with multi-modal fts. 84% accuracy (AUC$=0.90$) incl cognitive scores.
Fraser et al. [87]	Automatic detection of MCI based on topic modelling and information content in Swedish and English.	In-domain[11]: 67 pps, Gothenburg ds, Sweden. Out-of-domain[12]: 96 pps, Karolinska ds Sweden; 78 pps, Pitt, USA.	Cognitive ast. **Speech task:** picture description (Cookie Theft) spoken or written (Karolinska ds).	HC and MCI. Based on cognitive ast and clinical diagnosis.	Linguistic analysis for ft extraction, multilingual topic models for dimensionality reduction and ft seletion, ML for group classification.	Detection performance: 63% accuracy in English; 72% accuracy in Swedish. Based on information content.
Fraser et al. [9]	Automatic detection of AD based on linguistic (mostly) and acoustic fts from narrative speech.	473 samples[13]: 233 HC (151F), 240 AD (158F). Avg age: 65.2, 71.8; educ: 14.1, 12.5, MMSE: 29.1, 18.5.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Acoustic and linguistic analysis of speech for ft extraction, factor analysis for dimensionality reduction, ML for group classification.	Detection performance: 81.92% accuracy. Factors identified (4): semantic, acoustic, syntactic and information.
Gonzalez-Moreira et al. [89]	Automatic detection of mild dementia based on acoustic fts from narrative speech in Spanish.	20 pps: 10 HC (1F), 10 CI[14] (4F). Avg age 78.9, 80.3; educ 7.8, 4.	Cognitive ast. **Speech task:** read short text aloud ("The Grandfather Passage".	HC and CI. Based on cognitive ast (i.e. MEC scores).	Acoustic analysis for ft extraction of recorded speech, ML for group classification.	Detection performance: 85% accuracy based on four prosodic fts (articulation rate, mean syllables duration, F0 sd and mean).
Gosztolya et al. [63]	Automatic detection of AD and MCI, as well as discrimination between HC, MCI and AD based on acoustic (ASR) and linguistic speech fts.	75 pps: 25 HC, MCI, AD. Avg age: 70.72, 72.4, 73.96; educ: 12.08, 10.84, 10.7; MMSE: 29.24, 27.16, 23.92. 225 recordings.	Cognitive ast. **Speech task:** narrative of previous day, immediate and delayed recall of 2 short films.	HC, MCI and AD groups. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog).	Acoustic (ASR) and linguistic analysis for ft extraction, ML for group classification.	Detection performance: 86% HC-AD, 80% HC-MCI, 81.3% HC-CI accuracy. Discrimination performance: 66.7% (morphologic and acoustic).
Guinn et al. [98]	Automatic detection of AD based on linguistic fts from dialogue transcripts.	56 pps from CCC[15]: 28 nonAD, 28 AD. Multiple transcripts per pp: 204 nonAD, 77 AD.	**Speech task:** conversational interview about pp's chronic condition and their experience in healthcare.	HC (nonAD: patients with chronic conditions unrelated to AD) and AD. Based on clinical diagnosis.	Linguistic analysis for ft extraction of dialogue transcripts (syntax, semantics, pragmatics), ML for gorup classification.	Detection performance: $Prec_{AD} = 80.8\%$, $recall_{AD} = 0.75\%$; $Prec_{nonAD} = 79.3\%$, $recall_{nonAD} = 82.1\%$.

[8] Mini-Examen Cognoscitivo (MEC) is the Spanish adaptation of the MMSE[34].

[9] In particular, a Spanish version of "The Grandfather Passage"Darley et al. [48].

[10] Short texts obtained from the International Reading Speed Texts (IReST).

[11] Data including MCI pps: 67 Gothenburg. 36 HC (23F), 31 MCI (16F). Avg age: 67.9, 70.1; educ: 13.1, 14.1; MMSE: 29.6, 28.2.

[12] Data not including MCI pps: 96 Karolinska (52F) and 78 Pitt (48): all HC. Avg age: 57.2, 63.9; educ: 13, 13.9; MMSE: N/A, 29.1.

[13] Pitt: repeated samples from 264 participants (97 HC, 167 AD).

[14] This category is named MD (mild dementia) in study.

[15] Carolina Conversations Collection (CCC) [54] is conversational corpus consisting of conversations about health and healthcare gathered longitudinally with people with different chronic conditions, including AD.

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Guo et al. [92]	Automatic detection of AD based on linguistic fts from narrative speech.	268 pps[16]: 99 HC (58F), 169 AD (114F). Avg age: 61.3, 71; educ: 13.3, 11.8; MMSE: 27.9, 18.7.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Linguistic analysis for ft extraction (phonetics, semantics, syntax, pragmatics) including perplexity, ML for group classification.	Detection performance: 85.4% accuracy including perplexity fts derived from language models.
Haider et al. [11]	Automatic detection of AD based on standardised acoustic ft sets extracted from narrative speech.	164 pps (Pitt): 82 HC (46F), AD (46F). Aged 50-80, mostly 65-75 (BWGB); educ over 7 years: MMSE over 10.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HCand AD. Based on cognitive ast (i.e. MMSE).	Acoustic analysis for ft extraction from the recorded narratives, ADR[17] for ft selection and representation, ML for group classification.	Detection performance: 71.34% accuracy with one ft set; 78-80% accuaracy with "hard fusion" of all ft sets.
Kato et al. [64]	Discrimination between HC, MCI and AD with a two-phase system based on speech fts cerebral blood flow.	48 pps: 20 HC (13F), 19 MCI (13F), 9 AD (4F). Age range: 64-92 years old. Other demographics unreported.	Cognitive ast. **Speech task:** topics hometown and childhood, HDS-R, memory tasks. Simultaneous fNIRS[18].	HC, MCI and AD. Based on cognitive ast. (i.e. CDR = 0, 0.5 or 1).	Multivariate statistics to generate cognitive rating (SPCIR) based on prosody, ML for group classification in two phases: fNIRS and prosody.	Discrimination performance: 85.4% overall accuracy. 32% MCI participants misclassified into HC group.
Khodabakhsh and Demiroğlu [84]	Automatic detection of AD based on acoustic fts from conversational speech.	54 pps: 27 HC (15F), 27 AD (10F). Age range: 60-80 years old. Other demographics unreported.	**Speech task:** pps were asked casual questions to elicit 10 min of spontaneous conversation.	HC and AD. Unreported criteria.	Voice activity detection (VAD) and acoustic analysis for ft extraction from recordings, ML for group classification.	Detection performance: 79.2% with best pair of fts (log of voicing ratio + avg absolute delta pitch.
Konig et al. [102]	Discrimination between HC, MCI and AD based on automatically extracted speech fts across different tasks.	Dem@care: 64 pps. 15 HC (9F), 23 HC (12F), 26 AD (13F). Avg age 72,73,80; educ:[19]: uni, col, hs; MMSE: 29,26,19.	Cognitive ast. **Speech task:** countdown, picture description, repetition, SVF (animals).	HC, MCI and AD. Based on subjective memory complain (HC), Petersen criteria (MCI), NINCDS-ADRDA (AD)	Acoustic analysis for ft extraction from speech recordings, statistics for ft selection, ML for group classification.	Detection performance: $EER_{HC-MCI} = 21\%$, $EER_{HC-AD} = 13\%$, $EER_{MCI-AD} = 20\%$. Equal ss-sp: 79%, 87%, 80%
Lopez-de Ipiña et al. [80]	Discrimination between HC and stages of AD based on acoustic fts, incl. emotional response (pilot study).	10 pps (AZTITXIKI[20]): 5 HC/AD (2F). AD: 1ES, 2SS, 2AS. Age label: middle (HC), elderly (HC and AD).	**Speech task:** telling pleasant stories, recounting pleasant feelings, conversational interaction.	HCand ES, IS, AS, which stand for early, intermediate and advanced AD. Unreported criteria.	Acoustic analysis and emotional response analysis (ERA) for ft extraction, ML for group classification.	Discrimination performance: 93.79% accuracy with speech and emotional fts. Unclear whether this result is on 4 groups or 2.
Lopez-de Ipiña et al. [79]	Discrimination between HC and stages of AD based on acoustic fts, incl. emotional response.	40 pps (AZTIAHORE[21]. 20 HC (10F), 20 AD (12F). AD: 4ES, 10SS, 6AS. Age range: 20-98 HC, 68-98 AD. Others unreported.	**Speech task:** telling pleasant stories, recounting pleasant feelings, conversational interaction.	HC[22] and ES, IS, AS, which stand for early, intermediate and advanced AD. Unreported criteria.	Acoustic analysis and fractal dimension for ft extraction, incl. emotional response, ML for group classification.	Discrimination performance: accuracy reported per class, avg. 96.89%. Overall performance unclear.
Lundholm Fors et al. [59]	Discrimination between HC, SCI and MCI based on syntactic fts extracted from narrative speech.	Gothemburg[23]: 90pps. 36 HC (23F), 23 SCI (14F), 31 MCI (16F). Avg age: 67.9, 66.3, 70.1; educ: 13.2, 16.1, 14.1; MMSE: 29.6, 29.5, 28.2.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC, SCI and MCI. Based on clinical diagnosis.	Linguistic analysis for syntactic ft extraction, statistical analysis for feature selection, and ML for group classification.	Discrmination performance (binary detection): HC-MCI: $F_1 = 0.68$, HC-SCI: $F_1 = 0.54$, SCI-MCI: $F_1 = 0.66$

[16] subset from Pitt with multiple samples per participant. This study used 498 samples: 242 HC, 256 AD.
[17] Active Data Representation: novel method presented in this paper.
[18] Functional near-infrared spectroscopy: measures cortical brain activity by monitoring changes of oxy/deoxygenated hemoglobin concentration.
[19] Mode, that is, most frequent educational category.
[20] Subset of AZTIAHORE, which is, in turn, a subset of AZTIAHO.
[21] which is, in turn, a subset of AZTIAHO
[22] Group annotated as CR (control group) in the paper, equated to HC for the purpose of this review.
[23] Gothenburgh MCI data set Wallin et al. [53]

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Luz [6]	Automatic detection of AD based on acoustic fts extracted directly from voice recordings of narrative speech.	398 recordings (Pitt): 184 HC, 214 AD. Other pp information unreported.	Clinical ast (i.e. medical, cognitive). **Speech task:** narrative picture description (Cookie Theft).	HC and AD. Based on cognitive ast (i.e. MMSE).	Acoustic analysis to extract paralingusitic fts directly from speech recordings, no ft selection, ML for group classification.	Detection performance: 68% accuracy (baseline classification with simple algorithms for voice activity detection and speech rate).
Luz et al. [10]	Automatic detection of AD based on dialogical, content-free fts extracted from transcripts.	38 pps (CCC, ibid.) 21 nonAD (12F), 17 AD (15F). Age over 65. Other demographics unreported.	**Speech task:** conversational interview about pp's chronic condition and their experience in healthcare.	HC (nonAD: patients with chronic conditions unrelated to AD) and AD. Based on clinical diagnosis.	Dialogue analysis for ft extraction from conversational transcripts, Markov chains for data representation, ML for group classification.	Detection performance: 86.5% accuracy with vocalisations and speech rate.
Martinez de Lizarduy et al. [60]	Automatic detection of MCI and AD based on acoustic fts with a novel decision support system (ALZUMERIC).	SVF ds: 62 HC (36F), 38 MCI (21F). Avg age: 56.73, 57.15. PD ds[24]: 12 HC, 6 AD. SS ds (AZTIAHORE): 20 HC (9F), 20 AD (12F).	**Speech tasks**: SVF (animals), picture description (PD), spontaneous speech (SS, see AZTIAHORE).	SVF: HCand MCI. PD: HC and AD. SS: HC and AD. Unreported criteria.	ALZUMERIC system: acoustic analysis for ft extraction from voice samples, automatic ft selection, ML for group classification.	Detection performance: HC-MCI (SVF): 80% accuracy. HC-AD: 94% (PD) and 95% accuracy.
Meilan et al. [100]	Automatic detection of AD based on temporal and acoustic speech fts.	66 pps: 36 HC (80%F) 30 AD (68%FF). Avg age: 74.06, 78.66; educ: 7.30, 6.27; MMSE: 27.97, 18.07.	Cognitive ast. **Speech task:** reading familiar sentences on screen.	HC and AD. Based on NINCDS-ADRDA and cognitive ast (i.e. GDS, MMSE).	Acoustic and temporal speech analysis for ft extraction, ML for group classification.	Detection performance: 83.3% accuracy with speech fts such as voice breaks.
Mirheidari et al. [86]	Discrimination between ND and FMD[25], based on doctor-patient conversational fts.	30 pps: 15 FMD (9F), 15 ND[26] (8F). Avg age: 57.8, 63.73; MMSE: 28.87, 18.79.	Cognitive ast. **Speech task:** neurology consultation (conversation).	FMD and ND. Based on Schmidtke et al. [40] (FMD), Petersen and NINCDS-ADRDA (ND)	Automatic conversation analysis for ft extraction, ML for ft selection and group classification.	Discrimination performance: 97% classification accuracy between FMD and ND with top-10 fts.
Mirheidari et al. [85]	Discrimination between different conditions based on linguistic fts from conversational speech.	IMDB[27].: 50000 entries. Pitt: 473 narratives. Hallam: 45 conversations. IVA: 18 conversations. Seizure: 241 conversations.	**Written task:** movie feedback (IMDB). **Speech task:** picture description (Pitt), neurology consultation.	Pitt: HC, AD. Hallam: FMD, ND, DPD.[28]. IVA: FMD, ND, MCI. Unreported criteria.	ASR and linguistic analysis (vector representation) for ft extraction and selection, ML for group classification.	Discrimination performance: 65.8% (Hallam), 70% (IVA). Best binary: 93.7% FMD-DPD (Hallam), 100% FMD-ND (IVA).
Mirheidari et al. [8]	Automatic detection of ND based on speech fts, comparing neurologist-led with virtual-agent-led interactions (IVA).	HUM, 30pps: 15 FMD (9F), 15ND[29] (8F). Age: 57.8, 63.73; MMSE: 28.87, 18.79. IVA, 12 pps: 6 FMD (1F), 6 ND (3F). Avg age: 55.67, 65.83, ACE-R: 83.67, 59.57.	Cognitive ast. **Speech task:** neurology consultation (HUM) or avatar interaction (IVA).	FMD and ND. Based on Schmidtke et al. [40] (FMD), Petersen and NINCDS-ADRDA (ND)	Acoustic, linguistic and conversational analysis of recordings, comparing IVA-patient with neurologist-patient interactions. ML for group classification.	Detection performance: Neurologist-patient: 90.0% accuracy. IVA-patient: 90.9% accuracy.
Mirheidari et al. [12]	Discrimination between HC, FMD, MCI and ND based on acoustic and linguistic fts from IVA-led speech.	61 pps: 14 HC (8F), 10 FMD (6F), 18 MCI (12F), 19 ND (7F). Avg age: 69.4, 56.4, 62.2, 69.8.	Cognitive ast. **Speech task:** SVF (animals) and 10 question conversations. Both led by an IVA.	HC, FMD, MCI and ND. Based Schmidtke, (FMD), Petersen (MCI) and NINCDS-ADRDA (ND) criteria.	Acoustic, linguistic and conversational analysis of SVF answers and IVA-patient interactions. ML for ft selection and group classification.	4-way discrimination performance: 62% accuracy and ROC-AUC: 81.5% with top 22 fts (48% all).
Mirzaei et al. [49]	Discriminate between HC, MCI and AD based on acoustic fts from narrative speech.	48 pps: 16 HC, 16 MCI, 16 AD. Avg age: 72.7, 77.6, 77.9; MMSE: 28.6, 28.3, 22.4.	Cognitive ast **Speech task:** reading familiar sentences on screen.	HC, MCI, AD. Based on cognitive ast ((MMSE over 20 for inclusion).	Acoustic analysis for ft extraction, ML for ft selection (wrapper) and group classification.	Discrimination performance: 62% accuracy (three-way classification).

[24]Both are subsets from the Gipuzkoa-Alzheimer Project: http://www.cita-alzheimer.org/projects/gipuzkoa-alzheimer-project-basque-cohort

[25]ND: neurodegenerative disorder (e.g. AD). FMD: functional memory disorder)

[26]Heterogeneous ND group: 8 AD, 3 AD+vD, 2 MCI and 2 FTD (frontotemporal dementia).

[27]DS details. IMBD: text entries on movies feedback. Pitt (previously described). Hallam [86], IVA [55] and Seizure [125]

[28]DPD: depressive pseudo-dementia.

[29]Heterogeneous ND group: 8 AD, 3 AD+vD, 2 MCI and 2 FTD (frontotemporal dementia).

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Nasrolahzadeh et al. [65]	Discriminate between HC and three stages of AD based on higher-order spectral analysis of speech data.	60 pps[30]: 30 HC (15F), 6 FS (3F), 15 SS (6F), 9 TS (5F). Avg age: 75.6, 73.3, 70.6, 77.4; MMSE: 28.39, 27.5, 26.8, 23.8.	Clinical ast (i.e. cognitive, medical). **Speech task:** prompted to talk about personal stories and feelings.	HC and AD, subdivided in FS, SS and TS. AD subgroups diagnosed with NINCDS-ADRDA criteria.	Acoustic spectral analysis for nonlinear feature extraction, ML for for ft selection and group classification.	Discrimination performance: 97.71% accuracy with 4-way classifier based on higher-order spectral fts.
Orimaye et al. [75]	Automatic detection of AD based on linguistic fts extracted from narrative speech.	198 pps (Pitt): 99 HC and 99 AD. Avg age: 65.26, 70.45. Other demographics unreported.	**Speech task:** narrative picture description (Cookie Theft).	HC and ADUnreported criteria.	Linguistic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: $AUC = 0.93$ with 1000 top combined fts (syntacic, lexical, n-grams).
Prud'Hommeaux and Roark [73]	Detection of MCI based on automatic alignment scores between transcribed recall tasks and source narratives.	124 pps: 52 HC, 72 MCI. Demographics unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story (LM-WMS-III[31]).	HC (non-MCI) and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic analysis and text alignment for automatic scoring of recall tasks, ML for group classification.	Scoring performance: $F_1 = 0.791$ (alignment based). Detection performance: $AUC = 0.795\%$).
Prud'hommeaux and Roark [70]	Detection of MCI based on automatic alignment scores between transcribed recall tasks and source narratives.	235 pps: 163 HC, 72 MCI. Avg age: 87.3, 88.7. Avg educ: 15.1, 14.9. Gender unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story (LM-WMS-III).	HC (non-MCI) and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic analysis and graph-based text alignment for automatic scoring of recall tasks, ML for group classification.	Scoring performance: $F_1 = 0.891$. Detection performance: $AUC = 0.748$. Pitt:[32]: $AUC = 0.704$
Rentoumi et al. [88]	Automatic detection of AD based on linguistic fts extracted from written narrative data.	60 pps: 30 HC (14F), 30 AD (17F). Avg age: 68.03, 66.48; educ: 13.93, 12; MMSE: 28.26, 22.68.	Cognitive ast. **Written task:** narrative picture description (Cookie Theft).	HC (NC) and AD. Based on cognitive ast. ($MMSE_{AD} = 10 - 25$).	Computational linguistic analysis for text ft extraction (morphosyntactic, lexical). ML for group classification.	Detection performance: 80% accuracy. (88.5% with synthetically enlarged ds).
Roark et al. [78]	Automatic detection of MCI based on scores and speech fts from recorded cognitive tests.	74 pps: 37 HC, 37 MCI. Avg age: 88.8, 89.8; educ: 15.1, 14.5; MMSE: 28.2, 26.4. Gender unreported.	Cognitive ast. **Speech task:** immediate and delayed recall of the Anna Thomson story.	HC and MCI. Based on cognitive ast (i.e. CDR=0.5 for MCI diagnosis).	Linguistic and acoustic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: $AUC = 0.861$ (test scores and automatically derived speech and language fts).
Rochford et al. [76]	Automatic detection of CI based on pause distribution fts from narrative speech.	187 pps: 150 HC, 37 CI. Avg age (all): 72.44; MMSE: 27.68, 114 females. educ unreported.	Cognitive ast. Speech task: ready aloud a passage from a children's story.	HC and CI. Based on cognitive ast. ($MMSE_{HC} >= 27$, $MMSE_{CI} < 27$).	Linguistic and acoustic analysis for ft extraction, statistical analysis for ft selection, ML for group classification.	Detection performance: 68.66% acc ($AUC = 0.74$).
Sadeghian et al. [43]	Automatic detection of AD based on acoustic and linguistic fts from narrative speech, and with MMSE scores.	72 pps: 46 HC, 26 AD. Avg age: 71.43, 78.48; educ: 13.28, 13.81; MMSE: 28.70, 20.92.	Cognitive ast. **Speech task:** narrative picture description (new picture[33]).	HCand AD. Based on medical diagnosis.	Customised ASR for speech transcription, acoustic and linguistic analysis ft extraction, ML for ft selection and group classification.	Detection performance (acc): 88.3%: demogr.+acoustic. 91.7%: ASR linguistic. 94.4%: MMSE+manual ling
Satt et al. [61]	Discrimination between HC, MCI and AD based on acoustic fts (content-free) from voice recordings.	Dem@care: 89 pps. 19 HC (15F), 43 MCI (31F), 27 AD (24F). Avg age: 67, 73, 72.	**Speech task:** narrative picture description, sentence repetition (15), syllable repetition ("pa-ta-ka").	HC, MCI and AD. Based on medical diagnosis.	Speech segmentation (VAD), acoustic analysis for ft extraction, statistical analyis for ft selection, ML for group classification.	Discrimination performance: $EER_{HC-MCI+AD} = 18\%$ $EER_{HC-MCI} = 17\%$ $EER_{HC-AD} = 15.5\%$
Shinkawa et al. [71]	Automatic detection of MCI based on single modality and multimodal behavioural data (gait and speech).	34 pps: 19 HC (12F), 15 MCI (8F). Avg age: 71.63, 74.87; MMSE: 28.42, 25.33.	Clinical ast. **Speech task:** narrative picture description (Cookie Theft). Gait task: 5-meter walk.	HC and MCI. Based on Petersen criteria.	ASR for speech transcription, gait and linguistic analysis for ft extraction, statistic analysis and ML for ft selection and group classification.	Detection performance: Multimodal: 82.4% acc. Single modality: 76.6% acc each ($F_{1speech} = 0.733$, $F_{1gait} = 0.667$).

[30] 30 AD participants distributed in three levels: First Stage (FS), Second Stage (SS) and Third Stage (TS).)

[31] Logical Memory Test of the Weschler Memory Scale III [45]

[32] This is an attempt from to authors to apply their model on unseen data. Only results based on their graph-based method are reported here.

[33] https://acoustics.org/wp-content/uploads/2015/10/Sadeghian-Figure1b.jpg

Table 5: SPICMO (PICOS) table (ctd.)

Study	Study aim/design	Population	Interventions	Comparison groups	Methodology	Outcomes of interest
Tanaka et al. [94]	Automatic detection of CI based on audiovisual fts from dialogues with a computer avatar.	29 pps: 15 HC (4F), 14 CI[34] (4F). Avg age: 74.1, 76.3; educ: 10.5, 14.1; MMSE: 27.5, 21.4.	Cognitive ast. **Speech task:** 10-15 min interaction with an avatar (dialogue system).	HC and CI. Based on medical diagnosis.	Acoustic, linguistic and image analysis for ft extraction, statistics for ft selection, ML for group classification.	Detection performance: 83% unweighted acc. $AUC = 0.93$. $AUC_{ADonly} = 0.89$.
Thomas et al. [66]	Discrimination of different stages of CI based on linguistic fts from interviews.	95 pps (ACADIE[35]): 85 high, 73 low; 35 normal, 50 mild, 53 moderate, 20 severe.	Cognitive ast. **Speech task:** two interviews about donepezil, 12 weeks apart.	Two or four groups. Based on MMSE scores (0-15, 16-20, 21-24 and 25-30).	Linguistic analysis for ft extraction (lexical, n-gram), statistics for ft selection, ML for group classification.	Discrimination performance: 95% acc severe vs. normal. 69.6% moderate vs. mild. 50% 4-way classification.
Tóth et al. [99]	Automatic detection of MCI based on acoustic fts from narrative speech	84 pps: 36 HC (23F), 48 MCI (32F). Avg age: 64.13, 73.08; educ: 12.47, 11.82; MMSE: 29.17, 26.97;	Cognitive ast. **Speech task:** previous day, immediate/delayed recall of 2 short films.	HC and MCI. Based on cognitive ast (i.e. MMSE, CDT, ADAS-Cog)	Customised ASR and acoustic analysis for ft extraction, statistics for ft selection, ML for gorup classification.	Detection performance: 75% acc automatic procedure ($F_1 = 0.788$, $AUC = o.676$) .
Tröger et al. [77]	Automatic detection of AD based on acoustic fts from narrative speech.	Dem@Care: 115 pps. 47 HC (40F), 68 AD (38F). Avg age: 72.4, 78.9.	Cognitive ast. **Speech task:** two life events, previous day, picture description.	HC and AD[36] Based on medical diagnosis.	Acoustic signal processing for ft extraction, univariate ft selection, ML for group classification.	Detection performance: 89% acc relying solely on vocal fts (ASR and content-free).
Tröger et al. [96]	Discrimination between SCI, MCI and AD with a simulated telephone-based SVF test (feasibility study).	166 pps: 40 SCI (32F), 47 MCI (24F), 79 AD (40F). Avg age: 72.65, 76.59, 79; educ: 11.35, 10.81, 9.47; MMSE: 28.27, 26.02, 18.81;	Clinical ast. **Speech task:** recorded SVF answers (animals).	SCI, MCI and AD. Based on subjective reports (SCI), Petersen (MCI) and NINCDS-ADRDA (AD)	ASR, acoustic and linguistic analysis for ft extraction, ML for group classification.	ASR performance: VFER[37]$= 33.4\%$. $AUC = 0.855$
Weiner et al. [74]	Automatic detection of AD based on acoustic fts from conversational speech in German.	ISLE: 74 pps[38]. 98 samples: 80HC, 13 AACD. 5AD. Age range: 70-74 years old.	Clinical ast. **Speech task:** semi-standardized biographic interviews.	HC, AACD (ageing associated cognitive decline) and AD. Based on medical diagnosis.	Acoustic analysis for ft extraction (focus on pause patterns), ML for group classification.	Detection performance: 85.7% acc. $UAR = 0.66$ $F_{1_{HC}} = 0.92$,$F_{1_{AD}} = 0.80$,$F_{1_{AACD}} = 0.80$.
Weiner and Schultz [68]	Automatic prediction of the development of CI from conversational speech in German.	ISLE: 51 pp. 35 HC, 16 CI[39] (developed within three visits). Age range: 61-77 (1st-3rd visit).	Clinical ast. **Speech task:** semi-standardized biographic interviews.	*No change* (HC) and *Change* (CI). Based on whether they remained healthy or not.	VAD and acoustic analysis for ft extraction (focus on pause patterns), ML for group classification.	Prediction performance: 80.4% acc (overall). $Rc_{no-change} = 0.91$, $Rc_{change} = 0.56$
Yu et al. [97]	Automatic detection of CI based on speech fts collected through remote assessments.	ADCS[40]: 167 pps. Pre-processed 180 samples: 160 HC, 20 CI.	Clinical ast. **Speech task:** SVF and EBi/EBd[41] delivered by telephone system.	HC and CI. Based on longitudinal medical diagnosis.	Acoustic analysis for ft extraction (articulatory, phonemic), ML for ft selection and group classification.	Detection performance: Speech+scores: AUC$= 0.77$ Speech only: AUC$= 0.74$ Scores only: AUC$= 0.54$

[34]Heterogeneous CI group: 9 AD, 1 NPH (normal pressure hydrocephalus), 1 AD+NPH, 1 DBL (dementia with Lewy bodies)

[35]ACADIE study [126]. Pps are divided by MMSE in either 2 (high, low) or four groups (normal, mild, moderate, severe)

[36]Heterogeneous group: diagnosed with either AD or a form of mixed dementia (including AD).

[37]VFER: Verbal Fluency Error Rate

[38]subset from ILSE: Interdisciplinary Longitudinal Study on Adult Development and Aging [127].

[39]Heterogeneous group: AACD, MCD (mild cognitive disorder), AD, VAD (vascular dementia)

[40]ADCS: Alzheimer's Disease Cooperative Study. 4-year longitudinal data collection for home-based assessment.

[41]East Boston Immediate/Delayed: summarise a story immediately/delayed after listening to it).

2.1. Data details table

This table accounts for details of the datasets, as well as specific subsets, used in the reviewed studies. It is structured as follows:

- **Data set size**: number of participants or samples, including details on number of words, or number of hours recorded, when available.
- **Data type**, with two distinctions: a) writings, audio recordings and/or transcripts (abbreviated as per Table 3); b) monologues or dialogues. Monologues, in turn, are divided into spontaneous, narratives and answers to cognitive tests (most frequently fluency task), whilst dialogues are subdivided into three groups: structured, semi-structured and conversational. When available, information about transcription (i.e. software used, manual vs. automatic) is included.
- **Other modalities**: such as video, cognitive scores or motor measurements, when applicable ("NA" is written otherwise).
- **Data annotation**: group labels available in the data, corresponding with what was described in the comparison groups column of the SPICMO table. It includes groups' *n*, i.e. group size, as well as groups' *m*, i.e. number of speech/test samples per group, as sometimes these two figures differ (e.g. in longitudinal studies).
- **Data balance**: whether the dataset or subset used in the study is balanced in terms of age, gender and education. It accounts for dataset balance, within class balance and between class balance when applicable (see "keys to table interpretability", above, for acronyms). If a feature is not reported in the table, this is because it was not reported in the article.
- **Data availability**: whether the data used in the study is available to the wider research community.
- **Language**: language in which the dataset was collected, including country of origin, since many languages are spoken in more than one country.

Names of particular datasets are underlined (e.g. Pitt) The second table aims to provide the community with benchmark information about current databases and their availability, in order to highlight recurrent gaps that future research projects should target when designing their data collection procedures.

Table 6: Detailed Data information

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Beltrami et al. [91]	39 pps.	Narrative monologues. Rec. and manual tr. (Transcriber[42]).	Cognitive scores: MMSE, MoCA, GPCog, CDT, VF.	HC (CON) / MCI: $n = 20/19$.	CB. Text reports balanced demogr., but no figures.	Unreported.	Italian (Italy).
Ben Ammar and Ben Ayed [95]	Pitt: $m = 484$. No. of pps unreported.	Narrative monologues. Rec. and manual tr. (CHAT[43]).	Cognitive scores: MMSE.	HC / AD (Dementia): $m = 242/242$.	CB. Demogr. unreported.	Pitt avail. (DementiaBank). Enhanced unreported	English (US).
Bertola et al. [57]	100 pps.	Monologues: fluency task. Rec. (unclear).	Cognitive scores: MMSE, Katz, Lawton, SVF.	HC (NC), aMCI, a+mdMCI[44], AD: $n = 25$	no-CB/CB (ibid.). no-WCGB: HC. BCB: A, G, E.	Unreported.	Portuguese (Brazil).
Chien et al. [82]	60 pps: 30 HC *ad-hoc*, 30 AD (Mandarin_Lu). 3 tasks each: $m = 150$.	Narrative monologues and fluency task. Rec.	Cognitive scores: SVF.	HC (CH) / AD: $n = 30/30$; $m = 75/75$.	CB. Demogr. unreported.	Mandarin_Lu avail. (DementiaBank). HC unreported.	Chinese, Taiwanese (Taiwan).
Clark et al. [67]	158 pps.	Monologues: fluency task. Manual tr. (text file).	Cognitive scores: CDR, MMSE, SVF. NI meas.: MRI.	HC (CN)/MCI-non[45]/MCI-con: $n = 51/83/24$.	No-CB, no-GB. no-WCGB. BCB: no-A, no-G, E.	Unreported.	English (US).
D'Arcy et al. [93]	87 pps. 5 tasks each: $m = 435$.	Narrative monologues and fluency task. Rec. and manual tr.	Cognitive scores: MMSE, NART, Memory, SVF.	HC (MMSE $>$ 27) / CI (MMSE $\leqslant$ 27): $n = 50/37$.	No-CB, no-GB. WCGB: unreported. BCB: unreported.	Unreported.	English (Ireland).
Dos Santos et al. [90]	3 ds (HC, MCI): Pitt: 86 tr. S/N: 9.58, 10.97; W/S: 9.18, 10.33. Cs[46]: 40 tr. S/N: 30.80, 29.90; W/S: 12.17,13.03 ABCD: 85 tr. (46, 39); S/N: 5.23, 4.95; W/S: 11, 12.04.	Pitt: narrative monologues. Cs: narrative monologues. ABCD:narrative monologues (cognitive test). All ds: Rec. and manual tr.	Pitt: MMSE. Cs: NA. ABCD: NA.	HC / MCI: Pitt: $n = 43/43$ Cs: $n = 20/20$ ABCD: $n = 20/23$.	Pitt: CB, no-GB. No-WCGB. No-BCB(A,G). Cs: CB, no-GB. No-WCGB. No-BCB(A,G,E). ABCD: no-CB, GB. No-WCGB. No-BCB(A,G,E).	All ds: available as used in study upon request to authors.	Pitt: English (US). Cs: Portuguese (Brazil). ABCD: Portuguese (Brazil).
Duong et al. [69]	99 pps. 2 tasks each: $m = 198$.	Narrative monologues. Rec. and manual tr. (verbatim).	Cognitive scores: PENO[47], WMS, language, visual.	HC (NE) / AD: $n = 53/46$; $m = 106/92$.	no-CBn, no-GBn. no-WCGBn. BCBn: A, no-G, no-E.	Unreported.	French (Canada).
Egas López et al. [62]	Dementia ds: 75 pps. 3 tasks each: $m = 225$. BEA ds: $m = 44$.	Dementia: Narrative monologues. Rec. and ASR tr. (Kaldi[48]).	Cognitive scores: MMSE, ADASCog, CDT (Dementia).	Dementia: HC, MCI, AD, $n = 25$. BEA: unreported.	Dementia: CB, GB. WCGB unknown. BCB: A, G, E.	Dementia: unreported. BEA: unreported, but avail. online[49].	Dementia&BEA: Hungarian (Hungary).
Espinoza-Cuadros et al. [83]	19 pps.	Narrative monologues. Structured dialogues (test). Rec. and tr.	Cognitive scores: MEC (Spanish MMSE), HDS-R.	HC (non-MCI): $n = 11$; MCI: $n = 8$.	No-CB, no-GB. WCGB (HC only). BCB: A, G, E (unclear).	Unreported.	Spanish (Cuba).
Fraser et al. [7]	55 pps. 3 tasks each: $m = 165$.	Narrative monologues. Rec. and tr.	Eye-tracking. Comprehension questions.	HC / MCI: $n = 29/26$; $m = 87/78$.	no-CBn, no-GBn. no-WCGBn. BCBn: no-A, G, E.	Restricted upon request to authors.	Swedish (Sweden).
Fraser et al. [87]	Gothenburg, Got: 67pps Karolinska, Kar: 96 pps Pitt: 116 pps.	Narrative monologues. Rec. and tr. (Got, Pitt). Written (Kar).	Cognitive scores: MMSE.	Got / Kar / Pitt: HC: $n = 36/96/97$; MCI: n =31/NA/19	CB, GB: Pitt only. WCGB: Pitt only. BCB: A,G (Pitt), E(all)	Got & Kar: unreported. Pitt: avail. (DementiaBank)	Got & Kar: Swedish. Pitt: Eng (US)

[42] http://trans.sourceforge.net

[43] CHAT protocol: Codes for the Human Analysis of Transcripts [128].

[44] aMCI: amnesic single-domain; a+mdMCI: amnesic multiple-domain. Class-balance depends on whether they are considered 1 or 2 groups.

[45] MCI-non (non converters) and MCI-con (converters) refer to whether MCI pps converted to AD or not over a 4-year follow-up.

[46] Cs: retellings of Cinderella Story [129].

[47] PENO is a cognitive battery in French (Joanette et al., 1995). Two pictures, "Bank robbery" and "Car accident" were described in this study.

[48] Kaldi speech recognition toolkit [104].

[49] Available under an Academic-Non Commercial use licence: http://www.nytud.hu/adatb/bea/index.html

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Fraser et al. [9]	Pitt: 264 pps. Several visits: $m = 473$. W/S: 100.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 97/176$; $m = 233/240$.	no-CB*n*,CB*m*,no-GB*m*. no-WCGB*m*. BCB*m*: no-A, G, no-E.	Avail. (DementiaBank).	English (US)
Gonzalez-Moreira et al. [89]	20 pps.	Narrative monologues. Rec.	Cognitive scores: MEC (Spanish MMSE).	HC / CI (MD): $n = 10/10$.	CB, no-GB. no-WCGB. BCB: no-A, no-G, no-E.	Unreported.	Spanish (Cuba).
Gosztolya et al. [63]	Dementia ds: 75 pps. 3 tasks each: $m = 225$.	Narrative monologues. Rec. and phonetic ASR tr.	Cognitive scores: MMSE, ADASCog, CDT.	HC / MCI / AD: $n = 25/25/25$; $m = 75/75/75$	CB*n*, GB*n*. WCGB unreported. BCB*n*: A, G, E.	Unreported.	Hungarian (Hungary).
Guinn et al. [98]	CCC: 56 pp. Several visits: $m = 281$.	Conversational dialogues. Rec. and tr. (Ten Have[50]. Pps with 1+ tr. merged.	Video (not all pps).	HC (non-AD) / AD: $n = 28/28$; $m = 204/77$;	CB*n*, no-CB*m*. Demogr. unreported.	Unreported, but avail. on request (ibid.).	English (US).
Guo et al. [92]	Pitt: 268 pps. Several visits: $m = 498$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 99/169$; $m = 242/256$.	no-CB*n*, CB*m*, no-GB*n*. no-WCGB*n*. BCB*n*: no-A,no-G,no-E.	Avail. (DementiaBank).	English (US)
Haider et al. [11]	Pitt: 164 pps. Speech segments: $m = 4076$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD: $n = 82/82$; $m = 2033/2043$.	CB*n*,*m*, no-GB*n*. WCGB*n*. BCB*n*: A, G.	Avail. (DementiaBank).	English (US)
Kato et al. [64]	48 pps.	Narrative monologues. Rec.	Cognitive scores: CDR, HDS-R. NI meas.: fNIRS[51]	HC (NC)/MCI/AD: $n = 20/19/9$.	no-CB, no-GB. WCGB: AD only. BCB: no-A, no-G.	Unreported.	Japanese.
Khodabakhsh and Demiroğlu [84]	54 pps. 10 min conversation each.	Semi-structured dialogues. Rec.	NA	HC / AD (Patient): $n = 27/27$.	CB, no-GB. no-WCGB. BCB: no-G. Demogr. unreported.	Unreported.	Turkish.
Konig et al. [102]	64 pps. 4 tasks each.	Monologues: countdown, repetition, picture description, fluency task.	Cognitive scores: MMSE, VF, IADL.	HC[52] / MCI / AD: $n = 15/23/26$	no-CB, GB. WCGB: MCI, AD. BCB: no-A, no-G, no-E.	Unreported.	French (France).
Lopez-de Ipiña et al. [80]	AZTITXIKI: 10 pps. (subset of AZTIAHORE[53]).	Narrative monologues. Conversational dialogues. Rec.	Video.	HC (CR) / AD_{ES} / AD_{IS} / AD_{AS}: $n = 5/1/1/2$	no-CB, no-GB. no-WCGB. BCB: no-A, G.	Unreported.	Multilingual (ibid.).
Lopez-de Ipiña et al. [79]	AZTIAHORE (ibid.): 40 pps.	Narrative monologues. Conversational dialogues. Rec.	Video.	HC (CR) / AD_{ES} / AD_{IS} / AD_{AS}: $n = 10/4/10/6$	no-CB, no-GB. WCGB: HC only. BCB: A, no-G.	Unreported.	Multilingual (ibid.).
Lundholm Fors et al. [59]	Gothemburg: 90 pps.	Narrative monologues. Rec. and tr.	Cognitive scores: MMSE.	HC / SCI / MCI: $n = 36/23/31$.	no-CB, no-GB. WCGB: MCI only. BCB: A, no-G, no-E.	Unreported.	Swedish (Sweden).
Luz [6]	Pitt: Unreported No. pps. Several visits: $m = 398$.	Narrative monologues. Rec. and manual tr. (CHAT).	Cognitive scores: MMSE.	HC / AD (ATD): $m = 184/214$.	no-CB.*m*. Unreported CB*n*. Demogr. unreported.	Unreported, but avail, (DementiaBank).	English (US).
Luz et al. [10]	CCC: 38 pps. 17 non-AD and 21 AD.	Conversational dialogues. Rec. and tr. (Ten Have - ibid.)	Video (not all pps).	HC (non-AD) / AD: $n = 17/21$	no-CB, no-GB. no-WCGB. BCB: no-G. A, E unreported.	CCC avail. (ibid.) Study identifiers avail. on request to authors.	English (US).

[50] CCC was transcribed using the Ten Have method [130] and is available upon request through carolinaconversations.musc.edu/

[51] fNIRS: Functional near-infrared spectroscopy. It measures hemodynamic responses in the brain as a proxy to measure neuron behaviour.

[52] The HC group in this study is conformed by pps who did actually have memory concerns but did not meet any diagnostic criteria (i.e. SCI).

[53] In turn, a subset of AZTIAHO: 50HC, 9hours (80% after pre-processing) and 20AD, 60min (50%). AD group is conformed by three AD stages, namely, ES (early), IS (intermediate) and AS (advanced). Multilingual: English, French, Spanish, Catalan, Basque, Chinese, Arabian and Portuguese.

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Martinez de Lizarduy et al. [60]	AN: 100 pps. PD[54]: 18 pps. SS: 40 pps (AZTIAHORE subset).	AN: fluency task. PD: narrative monologues. SS: spontaneous monologues. Rec.	Video.	AN / PD / SS: HC: $n = 62/12/20$; MCI:n =38/NA/NA AD: =NA/6/20; PD & SS: unreported.	CB: SS only. No-GB. no-WCGB$_{AN}$. CBC$_{AN}$: A, no-G.	Unreported.	AN: unreported PD: unreported SS: multilingual (ibid.).
Meilan et al. [100]	66 pps.	Narrative monologues. Rec.	Cognitive scoreS: MMSE.	HC (control) / AD: $n = 36/30$	no-CB, no-GB. BCB: A, no-G, E. No-WCGB.	Unreported.	Spanish (Spain).
Mirheidari et al. [86]	30 pps: 15 ND, 15 FMD.	Semi-structured dialogues. Rec and manual (verbatim) and ASR tr.	Cognitive scores: MMSE.	FMD / ND: $n = 15/15$	CB, no-GB. WCGB: ND only. BCB: A, G (unclear).	Unreported.	English (UK).
Mirheidari et al. [85]	Pps/files/utt/h/MLU(s): Pitt: 255/473/473/8/61.1 Hallam: 117/45/8970/12/4.8 IVA: 40/18/785/3.25/14.9 Seizure: 597/241/28000/50/6.3	Pitt: narrative monologues. Rec. and tr. Hallam, IVA, Seizure: semi-structured dialogues. Rec and manual (verbatim) and ASR tr.	Cognitive scores: MMSE.	Pitt: HC, AC. Hallam: FMD, ND, DPD. IVA: FMD, MCI, ND. Seizure: different seizure diagoses.	n unreported. Demogr. unreported.	Pitt: avail (DementiaBank). Hallam: unreported. IVA: unreported. Seizure: unreported.	Pitt: English (US). Hallam: English (UK). IVA: English (UK). Seizure: English (UK).
Mirheidari et al. [8]	HUM: 30 pps. IVA: 12 pps.	HUM: structured dialogues. IVA: structured dialogues (with avatar). Rec and CA annotations.	Video (IVA only). Cognitive scores: MMSE, ACE-R.	FMD / ND: HUM: $n = 15/15$. IVA: $n = 6/6$.	HUM & IVA: CB, no-GB. WCGB: ND only. BCB: no-A, no-G.	Unreported.	English (UK).
Mirheidari et al. [12]	61 pps. 4.3h, 1944 utt, 85 spk (incl. chaperons), 8s MLU.	Monologues: fluency task. Structured dialogues (IVA). Rec. and ASR tr. (Kaldi).	Video. Cognitive scores: MMSE, ACE-R.	HC / FMD / MCI / ND:[55]. $n = 14/10/18/19$.	No-CB, no-GB. no-WCGB. BCB: no-A, no-G.	Unreported.	English (UK).
Mirzaei et al. [49]	48 pps. Avg samples length: 17.47 s.	Narrative monologues. Rec.	Cognitive scores: MMSE.	HC / MCI / AD: $n = 16/16/16$	CB. G & E unreported. BCB: A (MCI-AD only)	Unreported.	French (France).
Nasrolahzadeh et al. [65]	60 pps. 16h after pre-processing[56]. Segments (60s):$m = 960$	Spontaneous monologues. Rec.	Cognitive scores: MMSE, CDR.	HC/AD$_{FS}$/SS/TS: $n = 30/6/15/6$ $m = 720/70/110/60$	no-CB, no-GB. WCGB: HC only. BCB: no-A, no-G.	Unreported.	Persian (Iran).
Orimaye et al. [75]	Pitt: 198 pps. MLU: 4.03s HC, 2.65s AD.	Narrative monologues. Rec. and manual tr (CHAT)	Cognitive scores: MMSE.	HC / AD: $n = 99/99$	CB. CBC: no-A.	Study data avail on GitHub[57].	English (US).
Prud'Hommeaux and Roark [73]	124 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, WMS-III.	HC / MCI: $n = 52/72$	no-CB. Demogr. unreported.	Unreported.	English (US).
Prud'hommeaux and Roark [70]	235 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, WMS-III.	HC / MCI: $n = 163/72$	no-CB. BCB: A, E. Gender unreported.	Unreported.	English (US).
Rentoumi et al. [88]	60 pps.	Narrative monologues. Written.	Cognitive scores: MMSE.	HC (NC) / AD: $n = 30/30$	CB. GB. no-WCGB. BCB: A, G (unclear), E.	Unreported.	Greek (Greece).
Roark et al. [78]	74 pps. 2 tasks each.	Narrative monologues. Rec. and manual tr.	Cognitive scores: CDR, MMSE, WMS	HC / MCI: $n = 37/37$	CB. Demogr. unreported.	Unreported.	English (US).
Rochford et al. [76]	187 pps	Narrative monologues. Rec and manual tr.	Cognitive scores: MMSE.	HC / CI: $n = 150/37$	no-CB, no-GB. Class demogr. unreported.	Unreported.	English (Ireland).
Sadeghian et al. [43]	72 pps. Avg sample length: 75.1s (sd 61.0).	Narrative monologues. Rec and tr. (manual+ASR).	Cognitive scores: MMSE.	HC (NL) / AD: $n = 46/26$	no-CB. BCB: no-A, E.	Unreported.	English (US).

[54] AN and PD are the "animal naming" and "picture description" subsets from the Gipuzkoa-Alzheimer Project (PGA): http://www.cita-alzheimer.org/projects/gipuzkoa-alzheimer-project-basque-cohort

[55] HC: healthy control. FMD: functional memory disorder. MCI: mild cognitive impairment. ND: neurodegenerative disorder (i.e. AD).

[56] 32h recorded, 15 from HC and 17 from AD stages. After pre-processing 12h remain from HC, 4h from AD stages.

[57] https://github.com/soori1/ADresearch.

Table 6: Detailed Data information (ctd.)

Study	Data set/Subset size	Data type	Other modalities	Data annotation	Data balance	Data availability	Language
Satt et al. [61]	89 pps.	Narrative monologues. Sentence/syllable repetition. Rec.	NA	HC / MCI / AD: $n = 19/43/27$	no-CB, no-GB. No-WCGB. BCB: A (MCI-AD only), no-G.	Unreported.	Greek (Greece).
Shinkawa et al. [71]	34 pps.	Monologue narratives. Wizard[58] of Oz method. Rec.	Gait ast (positional 3D). MMSE scores.	HC / MCI: $n = 19/15$	no-CB, no-GB. WCGB: MCI only. BCB: A, no-G.	Unreported.	Japanese (Japan).
Tanaka et al. [94]	29 pps. Avg interaction: $m = 10 - 15min$.	Structured dialogues (avatar). Rec. and manual tr.	Eye-tracking. Video.	HC / AD: $n = 15/14$ $m = 7 - 3/8 - 22$	CB, no-GB. no-WCGB. BCB: A, G, no-E.	Unreported.	Japanese (Japan).
Thomas et al. [66]	ACADIE: 95 pps. $m = 158$	Conversational dialogues. Rec. and manual tr.	Cognitive scores: MMSE.	HC/ Mild/ Moderate/ Severe: $m = 35/50/53/20$	Unreported.	Unreported.	English (Canada).
Tóth et al. [99]	Dementia: 84 pps. 3 tasks each: $m = 252$. Dementia: unreported.	Narrative monologues. Rec. and phonetic ASR tr.	Cognitive scores: MMSE, ADASCog, CDT.	HC (NC) / MCI: $n = 36/48$	no-CB, no-GB. no-WCGB. BCB: no-A, G, E.	Dementia: unreported. BEA: unreported, but avail. online[59].	Hungarian (Hungary).
Tröger et al. [77]	115 pps. Avg sample length: 140s.	Narrative monologues and countdown task. Rec and ASR tr.	NA	HC / AD: $n = 47/68$	no-CB, no-GB. no-WCGB. CBC: no-A, no-G.	Unreported.	French (France).
Tröger et al. [96]	166 pps.	Monologues: fluency task.	Cognitive scores: MMSE, CDR.	SCI (SMC) / MCI / AD: $n = 40/47/79$	No-CB, no-GB. WCGB: MCI and AD. BCB: no-A, no-G, no-E.	Unreported.	French (France).
Weiner et al. [74]	ISLE: 74 pps. $m = 98$ (treated as n). 230h.	Semi-structured dialogues. Rec. and manual tr.	NA.	HC/ AACD[60]/ AD: $m = 80/13/5$	no-CB. Demogr. unreported.	Unreported.	German (Germany).
Weiner and Schultz [68]	ISLE: 23 pps. 112h $m = 51$ (treated as n).	Semi-structured dialogues. Rec. and manual tr.	NA	No-change/Change: $m = 35/16$[61]	no-CB. Demogr. unreported.	Unreported.	German (Germany).
Yu et al. [97]	167 pps. $m = 180$ (treated as n).	Narrative monologues and fluency task. Rec.	Cognitive scores: WMS-III,SVF,Trail	HC / CI[62]: $m = 160/20$	No-CB, unclear G. BCB: A, G, B	Unreported.	English (US).

[58] Wizard of Oz: experiment method by which human-computer interaction is examined. In this case the experimenter pretended to be the computer.

[59] Available under an Academic-Non Commercial use licence: http://www.nytud.hu/adatb/bea/index.html

[60] AACD: ageing-associated cognitive decline.

[61] HC who changed to AACD (ageing-associated cognitive decline), MCD (mild cognitive disorder), AD or VAD (vascular dementia)

[62] CI: cognitive impairment. Heterogeneous group including dementia, amnestic MCI single domain, amnestic MCI multiple domain). Recordings collected quarterly or annually (50-50%).

2.2. Methodology table

This table summarises the features and methods employed in the reviewed studies. It is structured as follows:

- **Pre-processing**: where available, this column describes the procedures undertaken on text and audio data as preparation steps for subsequent analysis. before. For text, this includes transcription (manual or ASR), tokenisation, removal of unanalysable events and *stopwords*, and so on. For audio, this includes background noise removal, normalisation, speaker diarisation.
- **Feature generation**: whether the features were generated from raw data through text analysis and/or through acoustic analysis, followed by more specific subcategories as per the taxonomy described in Table 1. When reported, this column also includes the paper's approach to reduce the extracted feature set, essentially either selection or extraction. On the one hand, *'filtering' selection* uses extrinsic criteria, such as information gain or, commonly, *p*-values (i.e. whether the differences between the experimental groups, e.g. AD and HC, for a particular feature are statistically significant or not); whereas *'wrapping' selection* uses a cross-validation model that searches through the power set of features. On the other hand, *feature extraction* entails creating a new reduced feature set by combining or transforming the original one with method such as PCA, LSA, clustering or ADR.
- **ML task/method**: supervised vs. unsupervised learning. Task: clustering, classification, regression. Method: clustering algorithm, classifiers and regression method as per Table 4. This column also includes information on the number of classes that the classifier outputs.
- **Evaluation technique**: describes four points, when available. First, the baseline against which the study results are compared (i.e. random guess, neuropsychological scores, different feature sets). Second, the performance metrics reported by the authors (i.e. *acc, F1, pc, rc, ss, sp, AUC, EER*, see Table 4). This will include information about different ASR precision measures, such as WER, where applicable. Third, the cross-validation technique used. Fourth, whether a test set held out, unused for model training, and its size.
- **Results**: numerical results of the selected performance metrics for the baseline and for the fitted model/s. When multiple metrics are reported, only summary metrics such as *EER, acc, F1* and *AUC* are included in this column.

Table 7: Methodology

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Beltrami et al. [91]	Processing unit: utt **Text:** manual tr. Paralinguistic annotation. **Audio:** VAD (ssvad[63]) and Kaldi[64]-ASR forced alignment	Filtering (selection): *p-values*. **Text-based:** lex diversity, PoS, lex density, syntactical (dependency). **Acoustic:** prosodic (temporal, F_0, energy), spectral.	Supervised learning. **Classification**: binary (HC-MCI) with *k*-NN ($k = 3$), LR and NN.	B/L: unreported. Metrics: *acc, pc, rc,* $F1$. CV: unreported. Hold-out set: 80/20%.	LR and NN performed best on "Picture" task: $acc = 76.9\%$, $pc = 0.727$, $rc = 0.842$ and $F1 = 0.781$.
Ben Ammar and Ben Ayed [95]	**Text:** ASR tr. **Audio:** removal of background noise and non-analysable[65] events.	Filtering: IG; Wrapping: *k*-NN, SVM. **Text-based:** lex diversity, lex density, syntactical (constituency), pragmatics (*UoL*).	Supervised learning. **Classification**: binary (HC-AD) with NN, SVM and DT.	B/L: no ft set reduction. Metrics: *acc*. CV: unreported. Hold-out set: unreported.	Best performance: $acc = 79\%$ SVM. Best ft set: *k*-NN ($acc = 69\%$ NN, 71% DT).
Bertola et al. [57]	**Text:** SVF word sequence $\rightarrow$ speech graph.	Filtering (selection): corr w/ cognitive ast. **Text-based:** syntactical (*SGA*).	Supervised learning. **Classification**: binary and 3-way with NB.	B/L: unreported. Metrics: *ss, sp, AUC*. CV: unreported. Hold-out set: unreported.	HC-MCI-AD, HC-MCI, MCI-AD: $AUC = 0.6 - 0.8$ HC-AD: $AUC > 0.8$ MCI subgroups: $AUC < 0.6$
Chien et al. [82]	Processing unit: syl **Text:** ASR tr, tokenization, pause annotation.	Filtering (selection): suitability, trainability, generalizability. DR: manual Feature Sequence. **Text-based:** syllable tokens, ASR-related (*FP, rep, dys*).	Supervised learning. **Classification**: binary (HC-AD) with bidirectional LSTM (RNN).	B/L: unreported. Metrics: *AUC*. CV: unreported. Hold-out set: 85/15%. Random shuffling.	$AUC = 0.954$.
Clark et al. [67]	Processing unit: word **Text:** fluency test manually transcribed for automatic scoring.	Wrapping (selection): RF (importance). **Text-based:** lexical (*BoW,n-grams*), syntactical (*SGA*), semantic (matrix decomposition: ICA), pragmatics (*coh*), fluency scores.	Supervised learning. **Classification**: binary (MCI: noncon) with ensemble RF, SVM, NB and MLP. Combined w/ LASSO	B/L: unreported. Metrics: *AUC*. CV: LOO. Hold-out set: unreported. Bootstrap.	$AUC = 0.872$ incl fluency scores. MRI enhances *sp* but not *ss*.
D'Arcy et al. [93]	**Text:** manual tr. **Audio:** removal of begin/end pauses> 250ms and visually inspected disturbances.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal), ASR-related (pauses patterns)	Supervised learning. **Classification**: binary (MMSE: low-high) with LDA.	B/L: unreported. Metrics: *acc*. CV: unreported. Hold-out set: unclear.	$acc = 76\%$ LDA. Avg vowel duration +17% in low MMSE group.
Dos Santos et al. [90]	**Text:** manual tr., utt segmentation, tokenization, removal of *stopwords*, punctuation, dysfluencies.	Wrapping (selection): majority vote in BoW, CN and CNE[66]. **Text-based:** lexical (*BoW*), syntactical adjacency network (*SGA*) enriched w/ semantic word embeddings.	Supervised learning. **Classification**: binary (HC-MCI) w/ GNB, *k*-NN, RF, SVM (linear and RBF). Multi-view and ensemble.	B/L: unreported. Metrics: *acc*. CV: 5-fold. Hold-out set: unclear.	Pitt: $acc = 65\%$ ensemble. Cinderella: $acc = 65\%$ SVM-RBF, CNE fts. ABCD: $acc = 75\%$ SVM-linear, BoW fts.
Duong et al. [69]	**Text:** manual tr (verbatim), discourse processing (multilayered cognitive model).	Ft set reduction: unreported. **Text-based:** lex diversity, lex density, syntactical (dependency, complexity), pragmatics (*UoL*).	Unsupervised learning. **Clustering**: Euclidean distance on discourse fts. **Factor analysis:** PCA.	B/L: unreported. Metrics: cluster *acc*. CV: N/A. Hold-out set: N/A.	Cluster composition: AD cluster: $acc = 61\%$ (sequence pic), $acc = 41\%$ (single pic)
Egas López et al. [62]	**Audio:** 25 ms signals, 10 ms time-shift. UBM[67] trained on BEA ds.	Extraction: i-vector[68] model fitted w/ UCM and *MFCCs*. **Acoustic:** spectral fts (*20 MFCCs*).	Supervised learning. **Classification**: binary (HC-MCI+AD), 3-way (HC-MCI-AD) w/ SVM.	B/L: unreported. Metrics: *acc, F1*. CV: 5-fold. Hold-out set: unreported.	$F1 = 0.792$, immediate recall task (binary). $acc = 56\%$, all utt (3-way).
Espinoza-Cuadros et al. [83]	Unreported.	Filtering (selection): *p-values*. **Acoustic:** prosodic (temporal: *SR, PR, PhR, AR*).	Supervised learning. **Classification**: binary (HC-MCI) w/ RF.	B/L: no ft set reduction. Metrics: *acc*. CV: LOO. Hold-out set: unreported.	$acc = 78.9\%$, RF (20 trees). Same *acc* w/ all fts and significant fts.

[63] VAD proposed by [131]

[64] http://kaldi.sourceforge.net/about.html

[65] Non-analysable events in this context refers to breaks, overlapping speech, coughing, laughter, short hard noises and the like.

[66] These are different feature spaces (BoW: Bag of Words; CN: Complex Networks; CNE: Complex Networks Enriched with word embeddings.

[67] UBM: Universal Background Model, trained to represent speaker-independent distribution of features [132]

[68] Dimensionality reduction method of the GMM supervector (Gaussian Mixture Model). It assumes each utt is produced by a different speaker

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Fraser et al. [7]	**Text:** manual tr. **Audio:** unreported. + Eye-movement + comprehension.	Ft set reduction: unreported. **Text-based:** lex diversity, lex density, *PoS*, syntactical (dependency). **Acoustic:** prosodic (temporal), ASR-rel. (*FP, dys*).	Supervised learning. **Classification**: binary (HC-MCI) w/ LR and RBF-SVM (Platt's[69]). Cascade: mode,task,session	B/L: train w/ cognitive scores. Metrics: *AUC, acc, ss, sp*. CV: LPO. Hold-out set: unreported.	B/L: $AUC = 0.75$, $acc = 65\%$. Best: $AUC = 0.88$, $acc = 83\%$, task level (both LR and SVM).
Fraser et al. [87]	**Text:** manual tr., removal of dysfluencies, laughter, PoS, lemmatization, extrat Ns and Vs.	Ft set reduction: unreported. **Text-based:** lex density, *n*-gram embeddings (*FastText*), topic modelling (cosine distance, topic frequency, words per topic).	Supervised learning. **Classification**: binary (HC-MCI; HC-AD) w/ linear SVM.	B/L: train w/o topic model fts. Metrics: *acc, ss, sp*. CV: LOO. Hold-out set: unreported.	Multilingual topic model: $acc = 63\%$ English (MCI); $acc = 72\%$ Swedish (MCI). $acc = 82\%$ English (AD).
Fraser et al. [9]	**Text:** word-level tr. and utt segmentation. Remove false starts and *FPs* (other *dys* remain). **Audio:** MP3 to mono WAV.	Filtering (selection): Pearson's corr. **Text-based:** *BoW*, lex diversity/density, *PoS*, syntactical (constituency), semantic (*PsyLing*), pragmatics (*UoL*). **Acoustic:** spectral (*MFCCs*).	Supervised learning. **Classification**: binary (HC-AD) w/ multilinear LR. + Factor analysis.	B/L: unreported. Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	$acc = 81.92\%$ w/ 35 top fts (drops w/ 50+). Four factors: semantic, acoustic, syntactic, information content.
Gonzalez-Moreira et al. [89]	**Audio:** bandpass filter, subband selection, temporal weight, subband corr, Gaussian filter, energy thresdold, F_0 detection.	Filtering (selection): *p-values*. **Acoustic:** automatic syllable nuclei detection to extract prosodic fts (temporal, F_0 and functionals in semitones).	Supervised learning. **Classification**: binary (HC-CI) w/ SVM.	B/L: unreported. Metrics: *acc, ss, sp*. CV: LOO. Hold-out set: unreported.	$acc = 85\%$, $ss = 81.8\%$ and $sp = 88.8\%$, w/ prosodic temporal fts and F_0.
Gosztolya et al. [63]	**Text:** phone-based ASR[70] tr., phonectic segmentation, time-aligned phoneme sequences.	Ft set reduction: unreported. **Text-based:** *PoS*, lex density, syntactical, semantic (topic words). **Acoustic:** phone based prosodic (temporal) and ASR-related (*FP, rep, hes*).	Supervised learning. **Classification**: binary (HC-MCI+AD) and 3-way (HC-MCI-AD) w/ SVM (SMO).	B/L: w/ demogr scores. Metrics: *acc, pc, rc, sp, F1, UAR*. CV: 5-fold. Hold-out set: unreported.	Binary: $UAR = 0.83$, $acc = 82.7\%$, $F1 = 86.3$ (B/L $acc = 68\%$). 3-way (only *acc*): $acc = 69.3$ (B/L 40%).
Guinn et al. [98]	**Text:** manual tr., subjects w/ multiple tr. conglomerated into one.	Filtering (selection): *p-values*. **Text-based:** *PoS*, lex diversity (*TTR, BI, HS*), syntactical (constituency) pragmatics (*UoL*).	Supervised learning. **Classification**: binary (HC-AD) w/ DT and NB.	B/L: unreported. Metrics: *pc, rc* (HC/AD). CV: LOO. Hold-out set: unreported.	$NB_{pc} = 79.3/80.8\%$, $NB_{rc} = 82.1/0.75\%$; $DT_{pc} = 67.9/67.9\%$, $NB_{rc} = 66.7/66.7\%$.
Guo et al. [92]	**Text:** manual tr., removal of annotation codes. Merge "Possible" and "Probable" AD into one AD group.	Filtering (selection): *AUC* (β). **Text-based:** *PoS*, lex diversity (perplexity), lex density, syntactical (constituency), pragmatics (*UoL*). **Acoustic:** prosodic (temporal, F_0), spectral (*MFCCs*), ASR-related (*FP*).	Supervised learning. **Classification**: binary (HC-AD) w/ LR, SVM, DT, RF, *k*-NN.	B/L: all 49 initial fts. Metrics: *acc*. CV: nested LOO. Hold-out set: unreported.	B/L: $acc = 74.8 - 80.7\%$ $acc = 76.8\%$ w/ unigram perplexity; $acc = 85.4\%$ w/ unigram perplexity + initial fts.
Haider et al. [11]	Create one AD group, matched for age and gender. **Audio:** VAD segmentation (energy thresghold= 65), 10s per segment, volume normalisation.	Filtering (selection): standard ft sets[71]. **Acoustic:** prosodic, spectral, vocal quality. Comprehensive ft sets: *emobase*, *ComParE*, *eGeMAPS*, *MRCG* functionals.	Supervised learning. **Classification**: binary (HC-AD) w/ DT, *k*-NN, LDA, RF and SVM.	B/L: random guess. Metrics: *acc, UAR*, confusion matrices. CV: LOO. Hold-out set: unreported.	B/L: $acc = 50.12\%$ $acc = 78.7\%$ w/ DT, hard fusion of ft sets and ADR[72].
Kato et al. [64]	**Audio:** phrase level segmentation, 23ms frames, Hamming window (1024 points). Voice extracted w/ short-time Fourier transform (every 11ms).	Extraction: PCA (+ stepwise regr). **Acoustic:** prosodic (F_0 and trajectories, energy), spectral (formant trajectories, *MFCCs*). + fNIRS[73] measures.	Supervised learning. **Classification**: binary, two-phased (first: HC-CI, second: MCI-AD) w/ NB. Empirical fts cut-off: 26/28.	B/L: unreported. Metrics: *acc, predictive value*. CV: LOO. Hold-out set: unreported.	$acc = 85.4$ w/ 26 cut-off, $acc = 83.3$ w/ 28 cut-off (this improves MCI classification from $acc = 94.7\%$ to $acc = 68.4\%$).

[69] Because SVM does not output probabilities directly.

[70] trained on BEA Hungarian Spoken Language Database Gósy [52].

[71] Standard feature sets available for openSMILE: https://www.audeering.com/opensmile/

[72] ADR: active data representation, novel method presented in this paper.

[73] fNIRS (functional near-infrared spectroscopy) measures cortical activity.

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Khodabakhsh and Demiroğlu [84]	**Audio:** VAD based on the distribution of the short-time frame energy (speech-silence). Automatic Turkish phoneme recogniser.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal, F_0, energy), spectral (formants).	Supervised learning. **Classification**: binary (HC-AD) w/ LDA, SVM and DT.	B/L: unreported. Metrics: *acc, TP, FA*, confusion matrices. CV: LOO. Hold-out set: unreported.	Best performance w/ SVM: $acc = 83\%$, $TP = 88.9\%$, $FA = 23.1\%$
Konig et al. [102]	**Audio:** VAD segmentation based on energy envelop and pitch contour (periodicity). Praat software.	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal, energy).	Supervised learning. **Classification**: binary (pairwise: HC, MCI, AD) w/ SVM.	B/L: unreported. Metrics: *EER*[74] *or where missclassification rates are equal.*. CV: random subsampling. Hold-out set: unreported.	$EER_{HC-MCI} = 21\%$ (*equal sp-ss* = 0.79). $EER_{HC-AD} = 13\%$ (0.87). $EER_{MCI-AD} = 20\%$ (0.80)
Lopez-de Ipiña et al. [80]	**Audio:** removal of background noise and non-analysable events, VAD segmentation.	Filtering: ft type; Wrapping: CV. **Acoustic:** prosodic (temporal, F_0, energy, *emo*), spectral (formants), vocal quality (*jitt, shimm, HNR*).	Supervised learning. **Classification**: binary (HC-AD) w/ polynomial SVM, MLP, *k*-NN, DT, NB.	B/L: no *emo* fts. Metrics: *acc, CER* (graph). CV: 10-fold. Hold-out set: unreported.	B/L: CER$= 17 - 25\%$ Performance: CER$= 2 - 20\%$ Best: $acc = 93.79\%$ w/ SVM and all *emo* fts.
Lopez-de Ipiña et al. [79]	**Audio:** removal of background noise and non-analysable events, VAD segmentation.	Selection: ft type and CV. **Acoustic:** ibid previous study. + ASR-related: Higuchi Fractal dimension (*FD*).	Supervised learning. **Classification**: binary (HC-AD) w/ MLP and *k*-NN.	B/L: no *FD* fts. Metrics: *acc, CER* (graph). CV: 10-fold. Hold-out set: unreported.	B/L: $CER \approx 14\%$. Best: $CER = 3.11\%$ ($acc = 96.89\%$) w/ MLP and comprehensive ft set.
Lundholm Fors et al. [59]	**Text:** manual tr. and dysfluency annotation.	Ft set reduction: unreported. **Text-based:** syntactical (constituency and dependency).	Supervised learning. **Classification**: binary (pairwise: HC, SCI, MCI) w/ RF.	B/L: unreported. Metrics: *F1*. CV: LOO. Hold-out set: unreported.	$F1_{HC-SCI} = 0.54$, $F1_{HC-MCI} = 0.68$, $F1_{SCI-MCI} = 0.66$.
Luz [6]	**Audio:** VAD segmentation based on amplitude (empirical threshold at -25dB). Syllable nuclei detection	Ft set reduction: N/A. **Acoustic:** prosodic (temporal: vocalisation events and speech rate).	Supervised learning. **Classification**: binary (HC-AD) w/ NB.	B/L: comparable paper. Metrics: *acc, F1, AUC*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 58.5\%$ Performance: $acc = 68\%$ ($AUC = 0.734\%$, $F1_{HC} = 0.70\%$, $F1_{AD} = 0.64\%$).
Luz et al. [10]	**Audio:** vocalisation graph generation[75] (VG). Syllable nuclei detection, speech rate normalisation.	Filtering (selection): with and w/o speech rate. **Acoustic:** prosodic (temporal: vocalisation events and speech rate), dialogue turn-taking patterns.	Supervised learning. **Classification**: binary (HC-AD) w/ additive LR. VGO (vocalisation), VGS (vocalisation + speech).	B/L: random guess. Metrics: *acc, pc, rc F1, AUC*. CV: LOO, 10-fold. Hold-out set: unreported.	B/L: $acc \approx 50\%$ VGO: $acc = 81.1\%$, $AUC = 0.798$. VGS: $acc = 86.6\%$, $AUC = 0.894$.
Martinez de Lizarduy et al. [60]	Matched: age and emotion. **Audio:** VAD segmentation in speech signal and dysfluencies (60s instances).	Filtering: *p*-values; Wrapping: CV. **Acoustic:** prosodic (temporal, energy, *loud*), spectral (formants, *MFCCs*), vocal quality (*jitt, shimm, HNR, NHR*). + ASR-related: Higuchi *FD*, entropy.	Supervised learning. **Classification**: binary (SVF: HC-MCI, PD: HC-AD, SS: HC-AD) w/ *k*-NN, SVM, MLP, CNN.	B/L: unreported. Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	SVF: $acc = 80\%$, PD: $acc = 94\%$, SS: $acc = 95\%$, w/ CNN.
Meilan et al. [100]	**Audio:** unreported.	Ft set reduction: unreported. **Acoustic:** prosodic (temporal, F_0, *loud*, energy), vocal quality (*jitt, shimm, HNR, NHR*).	Supervised learning. **Classification**: binary (HC-AD) w/ stepwise LDA.	B/L: unreported. Metrics: *acc*. CV: resubstitution. Hold-out set: unreported.	no-CV: $acc = 84.8\%$ (misclassified: 4 HC, 6 AD). CV: $acc = 83.3\%$ (misclassified: 4 HC, 7 AD).
Mirheidari et al. [86]	**Text:** ASR tr., diarization, conversion to XML, turn start time equated to previous turn end time.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity, semantics (*FW*, topic modelling). **Acoustic**: ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** binary (FMD-ND) w/ linear SVM, RF, AdaBoost, MLP, SGD.	B/L: no ft set reduction. Metrics: *acc*. CV: LOO. Hold-out set: unreported.	B/L: $acc = 93\%$ Top-10 fts: $acc = 97\%$ w/ SVM, AdaBoost and SGD.
Mirheidari et al. [85]	**Text:** ASR tr., diarization.	Ft set reduction: unreported. **Text-based:** *BoW*, neural word embeddings (*GloVe*: vector average/variance and cosine distance).	Supervised learning. **Classification:** binary and 3-way (FMD, DPD, MCI) w/ LR and CNN-LSTM	B/L: manual approach. Metrics: *acc, WER* (ASR). CV: 10-fold. Hold-out set: unreported.	Binary / 3-way. B/L:*acc*=50-81.25/66.5-70% LR: *acc*=62−100/65.8−70% CNN_LSTM: *acc*=62.3%

[74]EER: Equal Error Rate, the point at which false alarm rate equals misdetection rate. Also the point were specificity=sensitivity (specificity-sensitivity = 1- EER/100)

[75]Markov diagrams encoding conditional transition probabilities between vocalisation events and steady-state probabilities. Vocalisation events: patient/interviewer talk, joint talk, silence (pause and switching pause).

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Mirheidari et al. [8]	**Text:** manual and ASR tr., diarization. **Audio:** unreported.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity. **Acoustic**: prosodic (temporal, $F-0$), vocal quality (*jitt, shimm, HNR, NHR*), ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** binary (FMD-ND) w/ linear SVM.	B/L: no ft set reduction. Metrics: *acc WER/DER*. CV: LOO. Hold-out set: unreported.	B/L: $acc = 90.0\%$ (manual tr). Top-10 fts: $acc = 100\%$ (manual tr), $acc = 90\%$ (ASR).
Mirheidari et al. [12]	**Text:** manual and ASR tr., diarization.	Wrapping (selection): RFE. **Text-based:** *BoW*, lex diversity, *PCA*. **Acoustic**: prosodic (temporal, $F-0$), vocal quality (*jitt, shimm, HNR, NHR*), ASR (dialogue: *TT, dys*).	Supervised learning. **Classification:** 4-way and binary (HC, FMD, MCI, ND) w/ LR.	B/L: no ft set reduction. Metrics: *acc, AUC, WER/DER*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 48 - 85\%$. Top-22 fts: $acc = 62 - 94\%$ (lowest for 4-way). $AUC_{4-way} = 0.815$
Mirzaei et al. [49]	**Audio:** band-pass filter (30-100 Hz), speech segmentation (10ms instances).	Wrapping (selection): two-stage. **Acoustic**: prosodic (temporal, F_0), vocal quality (*jitt, shimm, HNR*), spectral (*MFCCs, FBEs*).	Supervised learning. **Classification:** binary (pairwise: HC, MCI, AD) w/ *k*-NN, linear SVM, DT.	B/L: no ft set reduction. Metrics: *acc*. CV: 8-fold. Hold-out set: unreported.	B/L: $acc = 32 - 36\%$. Selected fts: $acc = 59 - 62\%$ DT (60%) selects 3 fts only.
Nasrolahzadeh et al. [65]	**Audio:** removal of background noise and non-analysable events. Segmentation (60s instances).	Filtering (selection): IG. **Acoustic**: ASR (*entr*), spectral. Higher order spectral analysis (*HOS*): bispectrum estimation FFT and AR.	Supervised learning. **Classification:** 4-way (HC-FS-SS-TS) w/ *k*-NN, RBF-SVM, NB, DT.	B/L: comparable paper. Metrics: *acc, ss, sp* (class). CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 81 - 97.96\%$. FFT fts: $acc = 95.42\%$ DT. AR fts: $acc = 97.71\%$ *k*-NN
Orimaye et al. [75]	Pp selection (last visit). **Text:** manual tr.	Filtering (selection): *p*-values. **Text-based:** *BoW* (*n*-grams), syntatical (constituency, dependency), semantic (*FW*), pragmatics (*UoL: rep, dys*).	Supervised learning. **Classification:** binary (HC, AD) w/ SVM (SMO).	B/L: previous work. Metrics: *AUC*. CV: LPO. Hold-out set: unreported.	B/L: $AUC = 0.75$. Top-1000 fts: $AUC = 0.93$
Prud'Hommeaux and Roark [73]	**Text:** manual word-level tr., tokenisation, downcase. Removal of partial words, punctuation, fillers.	Ft set reduction: unreported.	Supervised learning. **Classification:** binary (HC-MCI) w/ SVM.	B/L: manual scores. Metrics: *AUC, pc, rc, F1*. CV: LPO. Hold-out set: alignment	B/L: $AUC = 0.822$ Training: $AUC = 0.795$ Weighting: $AUC = 0.784$ Intersection: $AUC = 0.767$
Prud'hommeaux and Roark [70]	**Text:** manual utt level tr., downcase. Removal of partial words, punctuation, fillers.	Ft set reduction: unreported. **Text-based:** automatic task scoring alignment based (retelling and phrase level) and graph based.	Supervised learning. **Classification:** binary (HC-MCI) w/ RBF-SVM.	B/L: manual scores, MMSE. Metrics: *AUC, pc, rc, F1*. CV: LPO. Hold-out set: alignment.	B/L: $AUC = 0.733 - 0.751$ Alignment: $AUC = 0.751$ Graph: $AUC = 0.748$ Pitt: $AUC = 0.832/0.823$
Rentoumi et al. [88]	**Text:** written data. Experiment A: $n = 60$ Experiment B: $n = 200$[76].	Ft set reduction: unreported. **Text-based:** lex diversty (*TTR, BI*), *PoS (word type freq)*, syntactical complexity (constituency).	Supervised learning. **Classification:** binary (HC-AD) w/ SVM (SMO) and NB.	B/L: ZeroR Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 0.50$ $NB_A = 78\%$, $NB_B = 85\%$; $SVM_A = 80\%$, $SVM_B = 88.5\%$.
Roark et al. [78]	**Text:** manual utt tr., manual syntactic annotation (Penn Tree-bank), automatic parsing (Charniak parser), manual and forced time-alignment.	Ft set reduction: unreported. **Text-based:** lex density, *PoS*, syntactical (constituency, dependency). **Acoustic:** prosodic (temporal), spectral (*MFCCs*).	Supervised learning. **Classification:** binary (HC-MCI) w/ SVMlight.	B/L: unreported Metrics: *AUC, corr*. CV: LPO. Hold-out set: unreported.	$corr = 0.87 - 0.96$ (manual-automatic fts) $AUC = 0.861$
Rochford et al. [76]	**Audio:** removal of background noise (high-pass filter) and breath. Full-wave signal rectification. Step segmentation.	Filtering (selection): *p*-values. **Acoustic:** distribution fts and prosodic temporal fts (conventional static and individual dynamic thresholds).	Supervised learning. **Classification:** binary (HC-CI) w/ LDA.	B/L: unreported Metrics: *acc, ss, sp, AUC*. CV: *k*-fold. Hold-out set: unreported.	Distribution: *acc*=68.66% (*AUC*=0.74) Static= 65.39% (0.69) Dynamic= 61.97% (0.58)
Sadeghian et al. [43]	**Text:** manual and ASR[77] tr. **Audio:** removal of begin/end pause and click. Signal normalisation. VAD for segmentation.	Wrapping (selection): best first greedy. **Text-based:** *LIWC*, *PoS*, lex diversity, lex density, syntactical (constituency). **Acoustic:** prosodic (temporal,F_0,*emo*).	Supervised learning. **Classification:** binary (HC-AD) w/ MLP.	B/L: MMSE scores Metrics: *acc, WER* (ASR). CV: 10-fold. Hold-out set: unreported.	B/L: $acc = 70.8\%$. Manual: $acc = 93.1\%$. ASR: $acc = 91.7\%$ Audio+demogr: *acc*=83.3%

[76] Synthetic samples created with SMOTE [133]

[77] Developed custom ASR with limited domain vocabulary and no requirement for real-time ASR. RNN GRU (Gated Recurrent Units) used for automatic punctuation.

Table 7: Methodology(ctd.)

Study	Pre-processing	Feature generation	ML task/method	Evaluation technique	Results
Satt et al. [61]	**Audio:** manual segmentation (silences above 60ms are pauses).	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal, energy).	Supervised learning. **Classification:** binary (HC-AD, HC-MCI, HC-*both*) w/ SVM.	B/L: unreported Metrics: *EER*. CV: 4-fold. Hold-out set: unreported.	$EER_{HC-AD} = 15.5\%$. $EER_{HC-MCI} = 17\%$. $EER_{HC-both} = 18\%$.
Shinkawa et al. [71]	**Text:** ASR tr., manual correction and annotation (fillers, false starts). **Audio:** microphone synchronisation.	Wrapping (selection): *ROC-AUC*. **Text-based:** *PoS*, lex diversity, semantic (cosine), syntactical (dependency). **Acoustic:** prosodic (temporal). + Gait fts.	Supervised learning. **Classification:** binary (HC-MCI) w/ linear SVM.	B/L: MMSE scores Metrics: *acc, ss, sp, F1*. CV: LOO. Hold-out set: unreported.	B/L:*acc*=76.5% (*F1*=0.667) Speech: *acc*=76.5% (0.733). Gait: *acc*=76.5% (0.667) Multimodal: *acc*=82.4% (0.813).
Tanaka et al. [94]	Avatar system: MMDAgent[78]. **Text:** manual utt tr and annotation, tokenisation. **Audio:** microphone gain set to 70dB. Separate video from audio.	Filtering (selection): *p*-values. **Text-based:** *PoS*, lex diversity (*TTR*), pragmatics (*UoL: hes*). **Acoustic:** prosodic (temporal, F_0, energy), vocal quality, dialogue (*TT*). + Image fts.	Supervised learning. **Classification:** binary (HC-AD) w/ linear SVM and LR.	B/L: unreported. Metrics: *AUC, acc*. CV: LOO. Hold-out set: unreported.	SVM: $AUC = 0.93$ ($acc = 83\%$); LR: $AUC = 0.91$ ($acc = 79\%$).
Thomas et al. [66]	**Text:** manual tr.	Ft set reduction: unreported. **Text-based:** *PoS*, lex diversity (*TTR, BI, HS*), semantic (clause-like unit, *n*-grams).	Supervised learning. **Classification:** binary (HC-severe/mild) and 4-way (HC, mild, moderate, severe) w/ CNG[79] and CWF.	B/L: ZeroR Metrics: *acc*. CV: unreported. Hold-out set: unreported.	HC-severe: B/L=63.6%, CWF=94.5%. HC-mild: B/L=58.8%, CWF=75.34-way: B/L=33.5%, CWF=50%.
Tóth et al. [99]	**Text:** ortographic and phonetic manual tr and annotation.	Filtering (selection): *p*-values. **Acoustic:** prosodic (temporal), ASR (*FP*). Automatic and manual extraction.	Supervised learning. **Classification:** binary (HC-MCI) w/ NB, RF and linear SVM (SMO).	B/L: manual, no ft set reduction Metrics: *acc,ss,sp,F1,AUC* CV: LOO. Hold-out set: unreported.	B/L: $F1 = 0.75$, $accwa = 71.4\%$ w/ SVM. Top-26, automatic: $F1 = 0.788$, $acc = 75\%$
Tröger et al. [77]	**Audio:** manual segmentation based on signal intensity, 25-28dB; silence length, 0.25-0.5s; minimum sound length, 0.1s.	Filtering (selection): mutual info. **Acoustic:** prosodic (temporal). Silence/sound segments, syllable information.	Supervised learning. **Classification:** binary (HC-AD) w/ SVM (RBF).	B/L: no ft set reduction Metrics: *acc*. CV: 10-fold. Hold-out set: unreported.	Top-23 fts: $acc = 89\%$.
Tröger et al. [96]	**Text:** manual and ASR tr. **Audio**: manual segmentation based on signal intensity.	Filtering (selection): clinical relevance. **Text-based:** *BoW, PoS*, lex diversity, semantic (neural word embeddings: distance). **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** binary (SCI-CI) w/ SVM.	B/L: no ft set reduction Metrics: *AUC, ss, sp, VFER* (ASR). CV: LOO. Hold-out set: unreported.	$VFER = 33.4\%$. Manual tr: $AUC = 0.852$. ASR tr: $AUC = 0.855$.
Weiner et al. [74]	**Text:** manual tr. Speaker segmentation (audio alignment). **Audio:** VAD segmentation (HMM recognizer).	Ft set reduction: unreported. **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** 3-way (HC-AACD[80]-AD) w/ LDA (SVD, no shrinkage).	B/L: unreported Metrics: *acc,UAR,pc,rc,F1*. CV: stratified 3-fold. Hold-out set: unreported.	$acc = 85.7\%$. $UAR = 0.66$ $F1_{HC}$=0.92, $F1_{AD}$=0.80, $F1_{AACD}$=0.33.
Weiner and Schultz [68]	**Text:** manual tr. Speaker segmentation (audio alignment). **Audio:** VAD segmentation (HMM recognizer).	Ft set reduction: unreported. **Acoustic:** prosodic (temporal).	Supervised learning. **Classification:** binary (no change-change[81]) w/ LDA (SVD, no shrinkage).	B/L: naively estimated *F1* Metrics: *acc, pc, rc, F1*. CV: stratified 6-fold. Hold-out set: unreported.	$Acc = 80.4\%$. No change / Change: $F1_{B/L}$=0.81, LDA=0.87 $F1_{B/L}$=0.48, LDA=0.64.
Yu et al. [97]	**Audio:** discard poor quality audio files, cross-session averaging.	Filtering (selection): Cohen's *d*. **Acoustic:** prosodic (temporal, F_0), spectral (formants)	Supervised learning. **Classification:** binary (HC-CI) w/ SVM and GC.	B/L: SVF score Metrics: *AUC*. CV: LPO. Hold-out set: yes (no %)	B/L: $AUC = 0.54$ GC, $AUC = 0.58$ SVM. GC: $AUC = 0.73$. SVM: $AUC = 0.75$.

[78] http://www.mmdagent.jp/

[79] CNG: Common *N*-grams approach. CWF: Common Word Frequencies.

[80] AACD: Age-associated cognitive decline.

[81] Intra-personal change measured by subtracting early speech vector from the later speech vector, and normalising resulting vector to unit length.

Clinical applicability

This table summarises our assessment of the potential implications and applications of findings of each reviewed paper as regards research and clinical use. The table is structured as follows:

- **Research implications**:
  - Research Novelty: whether at the time of publication the study described a new dataset, proposed a new set of features, implemented a new method or applied an existing one for a different task;
  - Study Replicability: *low*, *partial* or *full*, depending on how well the procedure is described and whether data or data identifiers are available). *Low* refers to cases where both data is unavailable and method description is incomplete or unsatisfactory; *partial* to cases where either is the case, and *full* when both data and methods are available and satisfactorily described.
  - Results generalisability: *low*, *moderate* and *high*, depending on whether the analysis is specific to the task, and/or there have been any extrinsic validation procedures and/or robust evaluation techniques are in place (i.e. train-test, CV, baseline). *Low* refers to cases where the analysis is indeed specific to the task, and therefore difficult to apply to other tasks (e.g. when relying heavily on content features). In *low* generalisability studies there are no extrinsic validation procedures (e.g. pilot in clinical settings) and the evaluation techniques are insufficient (e.g. CV is in place, but no train-test and/or appropriate baseline comparisons). The improvement of one of these features would bring the study up to *moderate*, and further improvements would make its generalisability *high*. Given the state of the field, no study is 100% generalisable, hence why we have used this terminology instead of the same we used for replicability. For generalisability to be *high*, most conditions need to be met except for the extrinsic validation, since it is still very uncommon in the field that studies are carried within a clinical setting.
- **Clinical potential**: external validation is outlined if present. That is, whether the procedure has actually been attempted in real life (*yes*); or is, at least, embedded in a device, or the experimental design envisions realistic clinical testing at some stage (*in-design*). This column also includes potential applications (i.e. early screening for new cases of SCI or similar, monitoring disease progression or supporting diagnosis of MCI and AD), potential outcomes for global health (i.e. language of study) and potential for the methodology to be remotely applicable (no, suggested potential, yes when tried or purposefully designed with that in mind).
- **Risk of bias**: Feature balance (*no/partial/yes*), suitable metrics (*yes/no*, i.e. whether metrics other than overall accuracy are reported when data are class-imbalanced), contextualized results (*yes/no*, i.e. whether an appropriate baseline is provided in order to put results into perspective), overfitting (*yes/no*, i.e. whether cross-validation and/or hold-out set procedures are implemented). With regards to sample size, we specify three ranges that ranges: $ds \leqslant 50$, $ds \leqslant 100$ and $ds > 100$.
- **Strengths/Limitations**: several characteristics are listed with a yes/no answer, "yes" indicating strength and "no" indicating limitation. These characteristics are:
  - spontaneous speech: speech data is naturally generated, generated in response to an open-answer question or a narrative task, or generated in response to a scripted cognitive task (i.e. verbal fluency or counting). Speech is considered spontaneous when it is natural and when its prompted by open-answered or narrative tasks. That is, for example, the Cookie Theft picture description would be spontaneous (although not natural), whereas reading sentences from a screen saying as many animals as possible within 60 seconds is not spontaneous (nor natural).
  - conversational speech: whether the study includes dialogue data or only monologue.
  - automation: the only characteristic that observes a 'middle' stage. Method automation can be labeled as *no*, when the only automated procedure is the ML task; *partial*, when aspects of the procedure other than the ML task, such as feature set reduction, are also automated; or *total*, when everything is automated including preprocessing (e.g. ASR is used for transcription).
  - content-independence: whether the model for feature generation relies heavily on content features of the data (e.g. lexical or high level *n*-gram are often closely related to the way in which spoken language was prompted).

* Transcription-free: text analysis usually requires transcripts. Whether manual or ASR, transcribing procedures entail many restrictions. Manual transcription is time-consuming, whereas ASR transcription have limited performance on impaired speech, and they need to be trained to a specific language, therefore adding an extra step to the method.

Table 8: Clinical applicability

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Beltrami et al. [91]	**Novelty:** preliminary results of new project (OPLON). **Replicability:** partial. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Italian sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** no. **Overfitting:** hold-out set, no CV. **Sample size:** $ds \leqslant 50$ ($n = 39$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial (manual tr). **Content-independence:** no. **Transcription-free:** no.
Ben Ammar and Ben Ayed [95]	**Novelty:** speech samples only. Compare three ft selection processes. **Replicability:** partial (unreported n). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** no hold-out set, no CV. **Sample size:** $ds > 100$ ($m = 484$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial (manual tr). **Content-independence:** no. **Transcription-free:** no.
Bertola et al. [57]	**Novelty:** graph analysis, MCI subtypes, 3-way classification. **Replicability:** partial (unclear performance metrics). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** disease progression. **Global Health:** Brazilian Portuguese words. **Remote application:** no.	**Feature balance:** yes[82]. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 100$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Chien et al. [82]	**Novelty:** ft selection based on suitability, trainability and generalizability. **Replicability:** partial (*ad hoc* fts & data). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Chinese syllables $\rightarrow$ generalisable to Taiwanese and Hakka. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** hold-out set, no CV. **Sample size:** $ds \leqslant 100$ ($n = 60$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Clark et al. [67]	**Novelty:** new fluency scores. Inclusion of MRI data. 4-year follow-up. Ensemble classifier. **Replicability:** full. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** disease progression. **Global Health:** US English words. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 158$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
D'Arcy et al. [93]	**Novelty:** ASR and prosodic fts (in 2008). **Replicability:** partial (incomplete data information and procedure). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Irish English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** no CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 87$)	**Spontaneous speech:** some. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Dos Santos et al. [90]	**Novelty:** complex networks enriched w/ word embeddings. Multi-view and ensemble classifiers. **Replicability:** full. **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English and Brazilian Portuguese sentences. **Remote application:** suggested potential.	**Feature balance:** *Pitt* and *Cs* CB. **Suitable metrics:** yes (CB $\rightarrow$ *acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 40 - 86$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Duong et al. [69]	**Novelty:** discourse analysis, cluster analysis. **Replicability:** partial (incomplete procedure). **Generalisability:** low (task-specific model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** French sentences. **Remote application:** no.	**Feature balance:** age only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** N/A. Reliability test. **Sample size:** $ds \leqslant 100$ ($n = 99$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Egas López et al. [62]	**Novelty:** i-vector approach, spectral fts only. **Replicability:** full **Generalisability:** high (2 ds, task-independent model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian sentences. **Remote application:** no.	**Feature balance:** yes[83]. **Suitable metrics:** yes (*acc, F1*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 75$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.

[82] aMCI: amnesic single-domain; a+mdMCI: amnesic multiple-domain. Class-balance depends on whether they are considered 1 or 2 groups.

[83] Class-balance depends on whether MCI and AD are considered 1 group (CI, better results) or 2 groups.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Espinoza-Cuadros et al. [83]	**Novelty:** prosodic fts only. Transcribed MEC. **Replicability:** full. **Generalisability:** moderate (task-independent model)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Cuban Spanish sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 19$)	**Spontaneous speech:** yes. **Conversational speech:** no[84]. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Fraser et al. [7]	**Novelty:** multimodal language data and eye-tracking. Cascaded classifiers. **Replicability:** full. **Generalisability:** moderate (different data types)	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** G & E only. **Suitable metrics:** yes (*AUC, acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 55$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Fraser et al. [87]	**Novelty:** topic models, multilingual word embeddings (English, Swedish). **Replicability:** full. **Generalisability:** high (different languages).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** multilingual model $\rightarrow$ higher performance. **Remote application:** no.	**Feature balance:** *Pitt* only. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 67 - 116$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Fraser et al. [9]	**Novelty:** comprehensive model (text-based and acoustic fts). **Replicability:** full. **Generalisability:** moderate (task-specific model).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English (*Pitt*). **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 264$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Gonzalez-Moreira et al. [89]	**Novelty:** Mild dementia. Specific tool and software[85]. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Cuban Spanish sentences. **Remote application:** no.	**Feature balance:** class only. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 20$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Gosztolya et al. [63]	**Novelty:** custom phone-based ASR, phonetic seg. **Replicability:** partial (incomplete procedure). **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian phonemes. **Remote application:** no.	**Feature balance:** yes[86]. **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 75$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** unclear. **Content-independence:** no. **Transcription-free:** no.
Guinn et al. [98]	**Novelty:** dialogue data, pragmatic fts. **Replicability:** partial (no pp IDs). **Generalisability:** moderate (representative data).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English dialogues. **Remote application:** no.	**Feature balance:** yes **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 56$)	**Spontaneous speech:** yes. **Conversational speech:** no[87]. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Guo et al. [92]	**Novelty:** comprehensive model, incl perplexity fts from LM. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** no **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 268$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

[84]Database contains conversational speech but it is not included in the analysis.

[85]DCGrab v3-0. Allows storing clinical and demographic data for each patient, as well as their voice.

[86]Class-balance depends on whether MCI and AD are considered 1 (CI, better performance) or 2 groups.

[87]Database contains conversational speech but specific dialogue features are not included in the analysis.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Haider et al. [11]	**Novelty:** comprehensive standard ft sets, enhanced data. ADR method. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** no.	**Feature balance:** yes **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 164$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Kato et al. [64]	**Novelty:** two-phase system w/ prosodic and physiological fts (cerebral blood flow). **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Japanese sentences. **Remote application:** no.	**Feature balance:** no **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 48$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Khodabakhsh and Demiroğlu [84]	**Novelty:** analyse ft pairs. Dialogue data. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Turkish dialogues. **Remote application:** no.	**Feature balance:** class only **Suitable metrics:** yes (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 54$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Konig et al. [102]	**Novelty:** dynamic time warping for ft extraction. **Replicability:** full. **Generalisability:** high (investigated w/ unseen data).	**External validation:** no. **Potential application:** disease progression. **Global Health:** French sentences. **Remote application:** suggested potential.	**Feature balance:** gender only **Suitable metrics:** yes (*EER*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 64$)	**Spontaneous speech:** no (SVF). **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lopez-de Ipiña et al. [80]	**Novelty:** preliminary results of new project (AZTIAHO). Emotional response fts. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc, CER*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 10$)	**Spontaneous speech:** yes. **Conversational speech:** no[88]. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lopez-de Ipiña et al. [79]	**Novelty:** emotional temperature and fractal dimension fts. **Replicability:** partial. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc, CER*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 40$)	**Spontaneous speech:** yes. **Conversational speech:** no[89]. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Lundholm Fors et al. [59]	**Novelty:** incl SCI pps, syntactic complexity only. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Swedish sentences. **Remote application:** no.	**Feature balance:** age only. **Suitable metrics:** yes (*F1*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 90$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Luz [6]	**Novelty:** vocalisation fts only. **Replicability:** low (unreported *n*). **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences (*Pitt*). **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($m = 398$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Luz et al. [10]	**Novelty:** turn-taking fts. Dialogue data. **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($m = 38$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.

[88] Database contains conversational speech but specific dialogue features are not included in the analysis.
[89] Database contains conversational speech but specific dialogue features are not included in the analysis.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Martinez de Lizarduy et al. [60]	**Novelty:** preliminary results of acoustic decision support system (ALZUMERIC). **Replicability:** partial. **Generalisability:** high.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** Multilingual model. **Remote application:** suggested potential.	**Feature balance:** not all three ds. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 40 - 100$)	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Meilan et al. [100]	**Novelty:** acoustic fts only. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Spanish sentences. **Remote application:** no.	**Feature balance:** age and educ only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 66$)	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Mirheidari et al. [86]	**Novelty:** doctor-patient consultation. Conversational fts. **Replicability:** full. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 30$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [85]	**Novelty:** doctor-patient consultation, human-robot interaction. Word-vector repr, conversational fts. Several ds. **Replicability:** low. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK/US English conversations. **Remote application:** suggested potential.	**Feature balance:** unreported. **Suitable metrics:** unclear (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** varied ($n = 40 - 255$)	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [8]	**Novelty:** compare doctor-patient consultation w/ human-robot interaction. Conversational analysis fts. **Replicability:** full. **Generalisability:** moderate.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 12 - 30$)	**Spontaneous speech:** no. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Mirheidari et al. [12]	**Novelty:** human-robot interaction for cognitive ast. 4-way classification. **Replicability:** full. **Generalisability:** low.	**External validation:** yes. **Potential application:** diagnosis support. **Global Health:** UK English conversations. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 12 - 30$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Mirzaei et al. [49]	**Novelty:** two-stage ft selection. Acoustic fts only. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** French sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 48$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Nasrolahzadeh et al. [65]	**Novelty:** HOS analysis of speech data. Best 4-way classifier (AD stages). **Replicability:** full. **Generalisability:** high.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Persian sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 60$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** yes.
Orimaye et al. [75]	**Novelty:** comprehensive linguistic fts, incl *n*-grams approach. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** suggested potential.	**Feature balance:** class only. **Suitable metrics:** no (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, unclear hold-out set. **Sample size:** $ds > 100$ ($n = 198$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Prud'Hommeaux and Roark [73]	**Novelty:** automatic word alignment for scoring recall task. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 124$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Prud'hommeaux and Roark [70]	**Novelty:** automatic graph-based word alignment for scoring recall task. **Replicability:** partial. **Generalisability:** high (translate to *Pitt*).	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 235$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** no. **Content-independence:** no. **Transcription-free:** no.
Rentoumi et al. [88]	**Novelty:** written data. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Greek sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*acc*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 60$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Roark et al. [78]	**Novelty:** combine speech fts and recall cognitive scores. Late onset MCI. **Replicability:** partial. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** yes. **Suitable metrics:** yes (*AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 74$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Rochford et al. [76]	**Novelty:** dynamic minimum pause threshold estimation (pause distribution). **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Irish English sentences. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 187$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.
Sadeghian et al. [43]	**Novelty:** compare combinations of manual, custom ASR and MMSE fts. **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** US English sentences. **Remote application:** no.	**Feature balance:** educ only. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 72$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Satt et al. [61]	**Novelty:** compare combinations of manual, custom ASR and MMSE fts. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Greek sentences and syllables. **Remote application:** suggested potential.	**Feature balance:** no. **Suitable metrics:** yes (*EER*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 89$).	**Spontaneous speech:** yes (some). **Conversational speech:** no. **Automation:** no. **Content-independence:** yes. **Transcription-free:** yes.
Shinkawa et al. [71]	**Novelty:** multimodal data (gait and speech). **Replicability:** full. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Japanese sentences. **Remote application:** suggested potential.	**Feature balance:** age only. **Suitable metrics:** yes (*acc, F1*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 34$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.
Tanaka et al. [94]	**Novelty:** human-robot interaction. Dialogue and image data (multimodal approach). **Replicability:** full. **Generalisability:** low.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** Japanese conversations. **Remote application:** yes.	**Feature balance:** yes. **Suitable metrics:** yes (*acc, AUC*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 50$ ($n = 29$).	**Spontaneous speech:** no. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** no. **Transcription-free:** no.

Table 8: Clinical applicability (ctd)

Study	Research implications	Clinical potential	Risk of bias	Strengths/Limitations
Thomas et al. [66]	**Novelty:** custom common *n*-grams algorithm. 4-way classification. **Replicability:** low. **Generalisability:** low.	**External validation:** no. **Potential application:** disease progression. **Global Health:** Canadian English conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** no CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 95$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** unclear. **Content-independence:** no. **Transcription-free:** no.
Tóth et al. [99]	**Novelty:** custom phone-based ASR, phonetic seg. Compare automatic and manual approach. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** Hungarian phonemes. **Remote application:** no.	**Feature balance:** gender & educ. **Suitable metrics:** yes (*acc, F1, AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 84$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** no.
Tröger et al. [77]	**Novelty:** infrastructure-free system, potentially remote and longitudinal within-subjects. Acoustic fts only. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** in-design. **Potential application:** diagnosis support. **Global Health:** French words/sentences. **Remote application:** yes (simulated).	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** no. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 115$).	**Spontaneous speech:** yes. **Conversational speech:** no. **Automation:** yes. **Content-independence:** yes. **Transcription-free:** no.
Tröger et al. [96]	**Novelty:** simulated telephone-based screening (SVF). **Replicability:** full. **Generalisability:** low.	**External validation:** in-design. **Potential application:** disease progression. **Global Health:** French words. **Remote application:** yes (simulated).	**Feature balance:** no. **Suitable metrics:** yes (*AUC*). **Contextualised results:** yes. **Overfitting:** CV, no hold-out set. **Sample size:** $ds > 100$ ($n = 166$).	**Spontaneous speech:** no. **Conversational speech:** no. **Automation:** yes. **Content-independence:** no. **Transcription-free:** no.
Weiner et al. [74]	**Novelty:** custom VAD algorithm. Longitudinal dialogue data[90]. 3-way classification. **Replicability:** low. **Generalisability:** low.	**External validation:** no. **Potential application:** diagnosis support. **Global Health:** German conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** yes (*acc, UAR*). **Contextualised results:** no. **Overfitting:** stratified CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 74$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Weiner and Schultz [68]	**Novelty:** prediction of within-subjects cognitive change. Custom VAD algorithm. Longitudinal dialogue data. **Replicability:** partial. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** German conversations. **Remote application:** no.	**Feature balance:** no. **Suitable metrics:** no (*acc*). **Contextualised results:** yes. **Overfitting:** stratified CV, no hold-out set. **Sample size:** $ds \leqslant 100$ ($n = 51$).	**Spontaneous speech:** yes. **Conversational speech:** yes. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** no.
Yu et al. [97]	**Novelty:** telephone-based cognitive ast. 4-year longitudinal collection[91]. Compare speech and cognitive scores. **Replicability:** full. **Generalisability:** moderate.	**External validation:** no. **Potential application:** disease progression. **Global Health:** US English sentences. **Remote application:** yes.	**Feature balance:** demogr, no class. **Suitable metrics:** yes (*UAC*). **Contextualised results:** yes. **Overfitting:** CV & hold-out set. **Sample size:** $ds > 100$ ($n = 165$).	**Spontaneous speech:** some. **Conversational speech:** no. **Automation:** partial. **Content-independence:** yes. **Transcription-free:** yes.

[90]Database contains longitudinal samples of conversational speech. However dialogue features are not included in the analysis, and samples by one pp are treated as different pps $\rightarrow$ subject dependence)

[91]Cross-observation averaging: discards longitudinal information, althought does not introduce subject dependence.